Rudolf Rosa

GPT et al.

Generating Texts with Transformer-Based Large Language Models

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Motivation: Transformer-based models

- Vaswani+, 2017: Attention is All you Need (NIPS)
  - State-of-the-art for most Natural Language Processing (NLP) tasks
  - Machine Translation human parity (*)
  - Passing the Turing Test (*)
- OpenAI: GPT-1 (2018) … GPT-4 (2023)
  - Large Language Models (LLM)
  - Generating texts, program code, processing data
  - Integration into tools (Copilot, Bing, Duolingo…)
  - Passing various highschool & university exams
  - Generating theatre play scripts (THEaiTRE)
- Multimodal models
  - Vision Transformers
  - GPT-4: input = text + image
  - PaLM-E: multimodal + embodiment

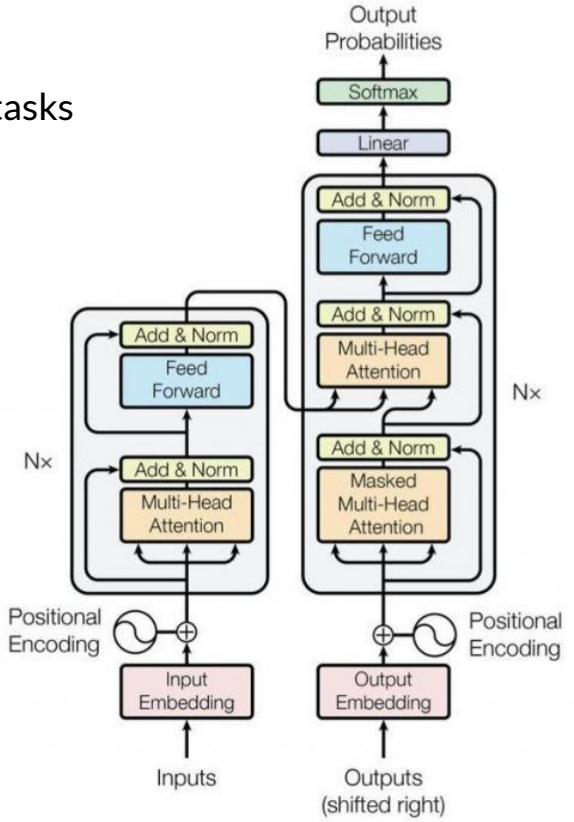


Figure 1: The Transformer - model architecture.

J. Alammar, https://jalammar.github.io/illustrated-gpt2/


Let’s watch a sample of “AI: When a Robot Writes a Play”!

DIRECTOR DANIEL HRBEK

AI: WHEN A ROBOT WRITES A PLAY

Outline

- Motivation
- Ground for Transformers
- Transformer architecture
- Models and applications

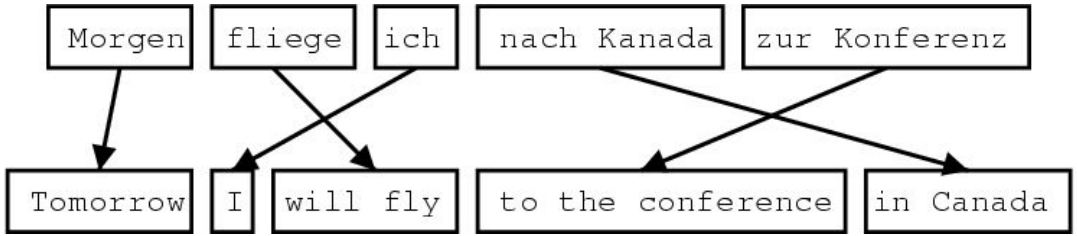
Running example: Machine Translation

- All human beings are born free and equal in dignity and rights.

- Všichni lidé rodí se svobodní a sobě rovní co do důstojnosti a práv.

Pre-neural Machine Translation

- Human translation
  - All human beings are born free and equal in dignity and rights.
  - Všichni lidé rodí se svobodní a sobě rovní co do důstojnosti a práv.
- Word-based
  - All human beings are born free and equal in dignity and rights .
  - Všechna lidská stvoření jsou zrozena svobodná a rovná v důstojnosti a právech .
- Phrase-based
  - All human beings are born free and equal in dignity and rights .
  - Všichni lidé rodí se svobodní a sobě rovní v důstojnosti a právech .
- Components
  - translation model = simple frequency tables → phrase translation suggestions
  - reordering model (word alignment)
  - language model → combines phrase translations into a fluent sentence



P. Koehn, https://www.statmt.org/wpt05/mt-shared-task/

Yet another example: Language Modelling

- Discriminative task: Which sentence is better (is more probable)?
  - All human beings are born free and equal in dignity and rights.
  - All human beings are born unmarried and equal in dignity and rights.
  - All humans beings are born free and equal in dignity and rights.
  - Free and equal in dignity and rights all human beings are born.

Yet another example: Language Modelling

- Discriminative task: Which sentence is better (is more probable)?
  - All human beings are born free and equal in dignity and rights.
  - All human beings are born unmarried and equal in dignity and rights.
  - All humans beings are born free and equal in dignity and rights.
  - Free and equal in dignity and rights all human beings are born.
- Generative task: Which word should follow (is more probable)?
  - I woke up in the morning and went to the…

Yet another example: Language Modelling

- Discriminative task: Which sentence is better (is more probable)?
  - All human beings are born free and equal in dignity and rights.
  - All human beings are born unmarried and equal in dignity and rights.
  - All humans beings are born free and equal in dignity and rights.
  - Free and equal in dignity and rights all human beings are born.
- Generative task: Which word should follow (is more probable)?
  - I woke up in the morning and went to the…
    - kitchen
    - bathroom
    - cinema
    - horse

Yet another example: Language Modelling

- Discriminative task: Which sentence is better (is more probable)?
  - All human beings are born free and equal in dignity and rights.
  - All human beings are born unmarried and equal in dignity and rights.
  - All humans beings are born free and equal in dignity and rights.
  - Free and equal in dignity and rights all human beings are born.
- Generative task: Which word should follow (is more probable)?
  - I woke up in the morning and went to the…
    - kitchen
    - bathroom
    - cinema
    - horse
- Semantic tasks:
  - What is the meaning of the sentence?
  - How similar are meanings of two given sentences?
  - …

Pre-neural Language Models

- Discrimination/generation
  - I woke up in the morning and went to the… kitchen/bathroom/cinema/horse
- N-gram language models (e.g. 4-grams)
  - How often are words A B C followed by word **D** (in first 100M lines of Wikipedia)?
    - "went to the **bathroom**" > "went to the **horse**"?
    - "went to the **cinema**" > "went to the **kitchen**"?
  - simple frequency tables, hard conditioning, limited context, no concept of word similarity
  - not usable for semantic tasks

Pre-neural Language Models

- Discrimination/generation
  - I woke up in the morning and went to the… kitchen/bathroom/cinema/horse
- N-gram language models (e.g. 4-grams)
  - How often are words A B C followed by word **D** (in first 100M lines of Wikipedia)?
    - "went to the **bathroom**" > "went to the **horse**"? 42 > 4
    - "went to the **cinema**" > "went to the **kitchen**"?
  - simple frequency tables, hard conditioning, limited context, no concept of word similarity
  - not usable for semantic tasks

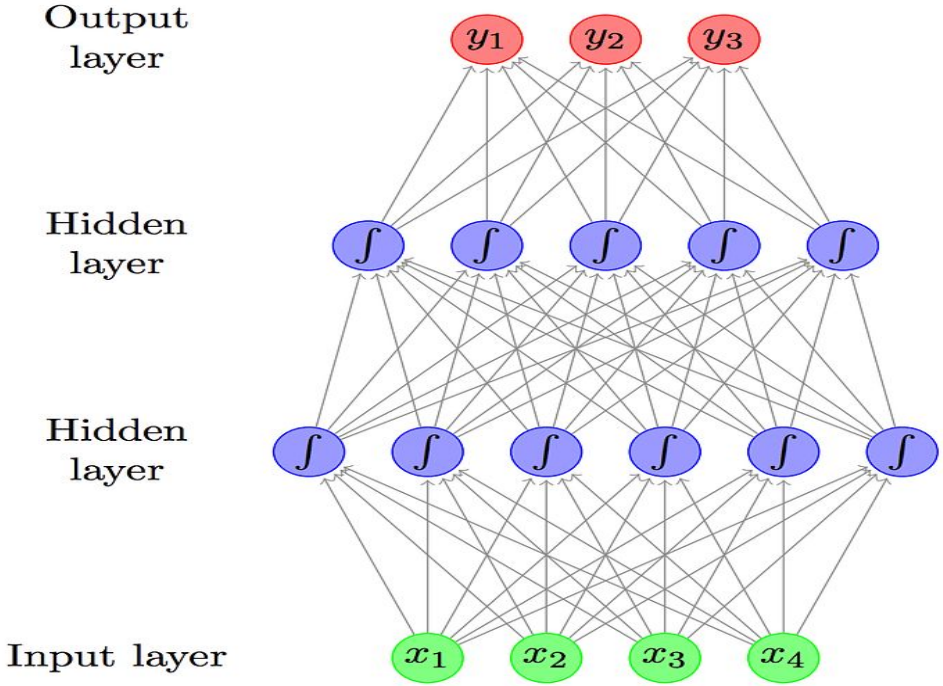
Pre-neural Language Models

- Discrimination/generation
  - I woke up in the morning and went to the… kitchen/bathroom/cinema/horse
- N-gram language models (e.g. 4-grams)
  - How often are words A B C followed by word **D** (in first 100M lines of Wikipedia)?
    - "went to the **bathroom**" > "went to the **horse**"? 42 > 4
    - "went to the **cinema**" > "went to the **kitchen**"? 25 > 14
  - simple frequency tables, hard conditioning, limited context, no concept of word similarity
  - not usable for semantic tasks

Ground for Transformers

Feed Forward Neural Network

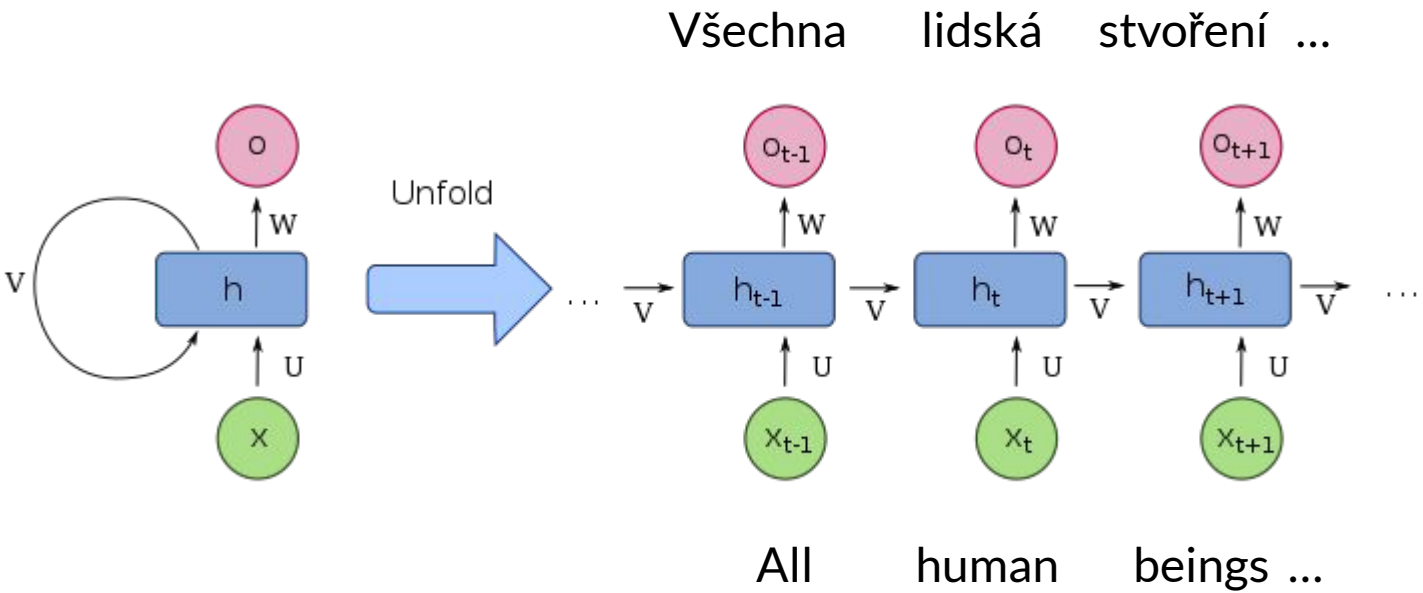
Všichni lidé se …



All human beings are …

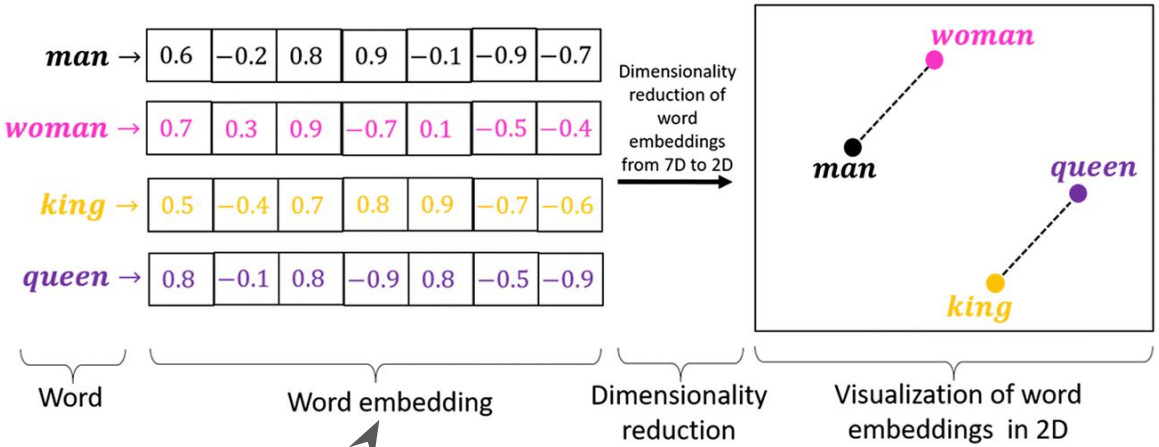
Gianp, *https://medium.com/mlearning-ai/feedforward-neural-networks-multi-layers-preceptors-mlps-1bea7ff11e07*

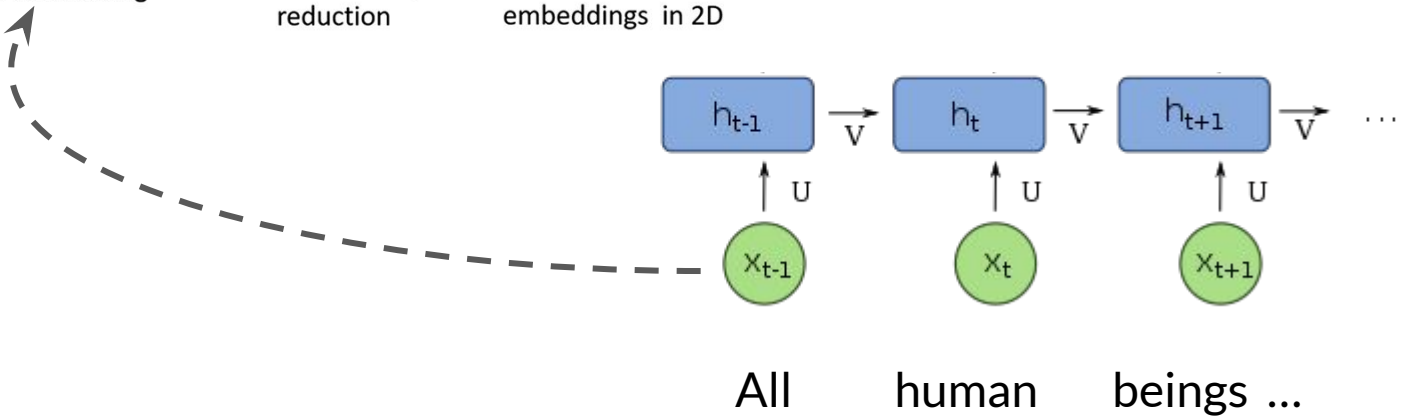
Recurrent Neural Network (RNN)



fdeloche, https://en.wikipedia.org/wiki/Recurrent_neural_network

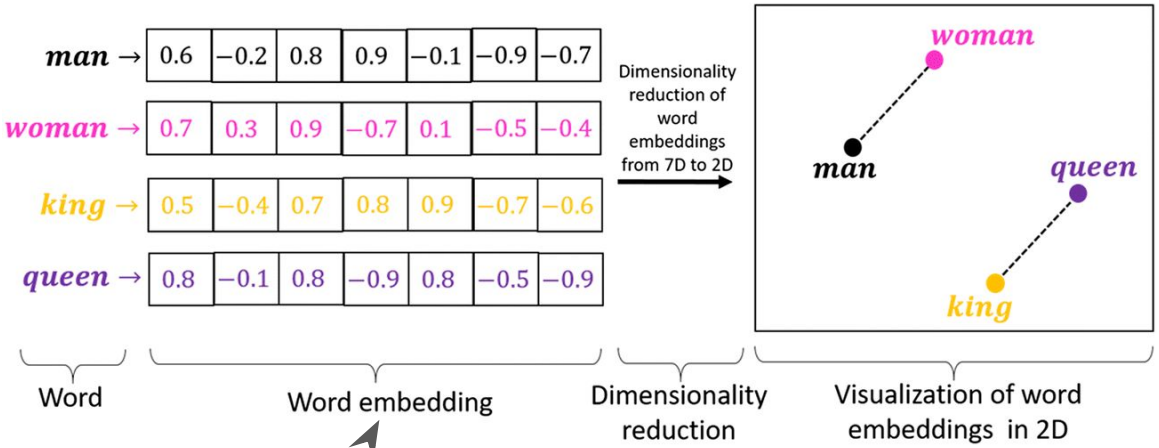
Word encoding: word embeddings (e.g. word2vec)





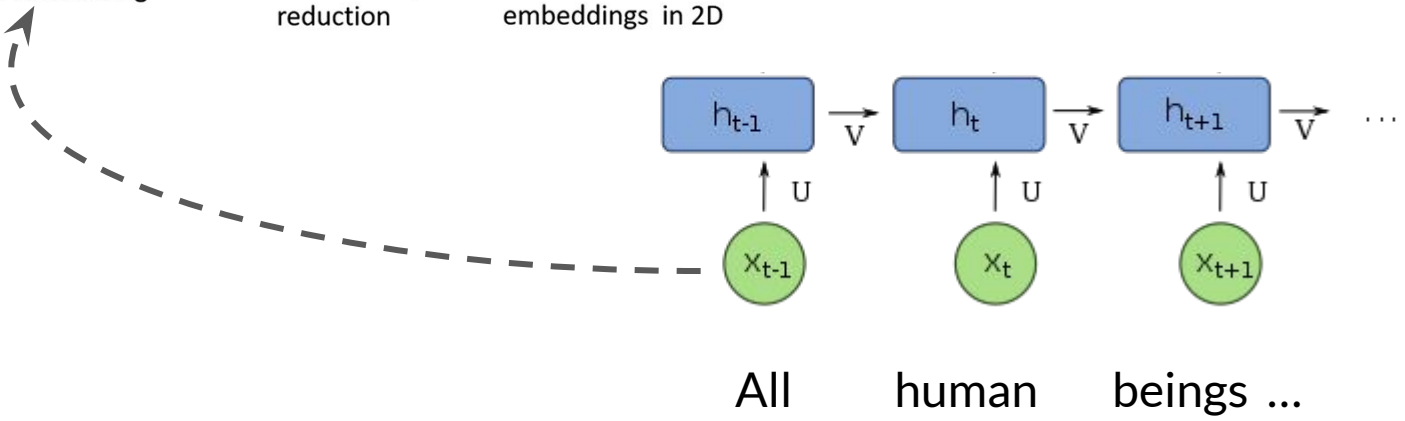
Siddharth M,
https://www.analyticsvidhya.com/blog/2021/07/feature-extraction-and-embeddings-in-nlp-a-beginners-guide-to-understand-natural-language-processing/

Word encoding: word embeddings (e.g. word2vec)



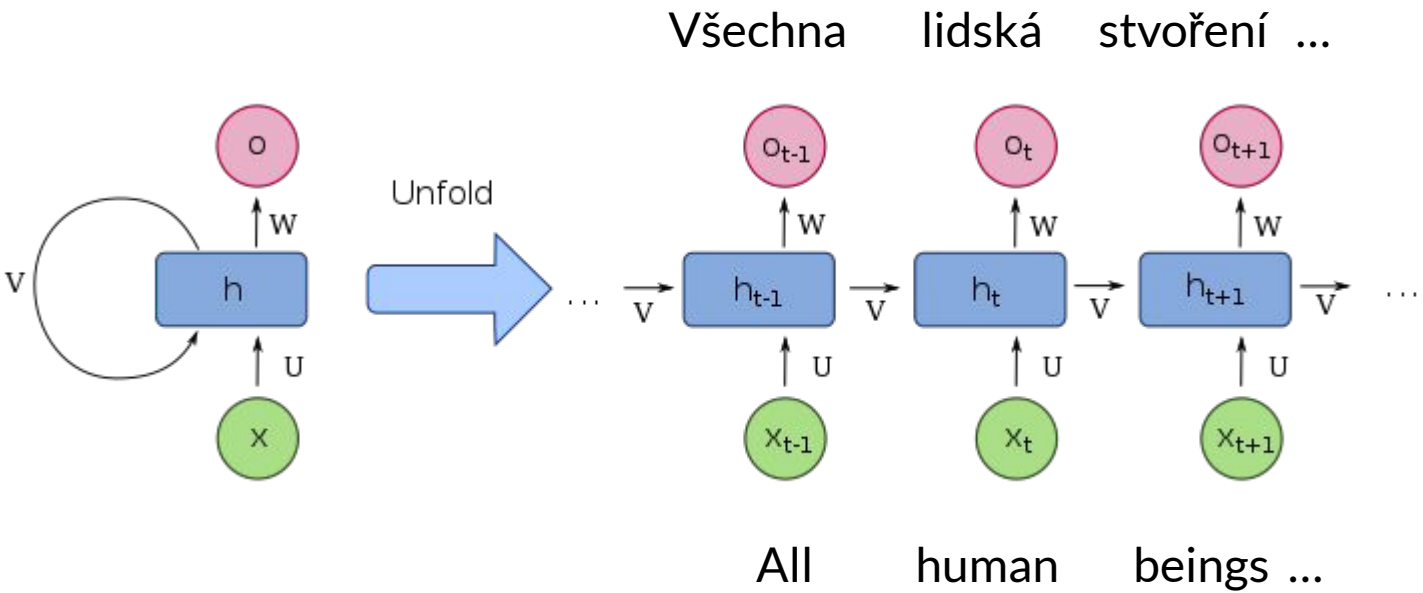
…and actually typically also split to subwords:

platypus → plat- ypu- s



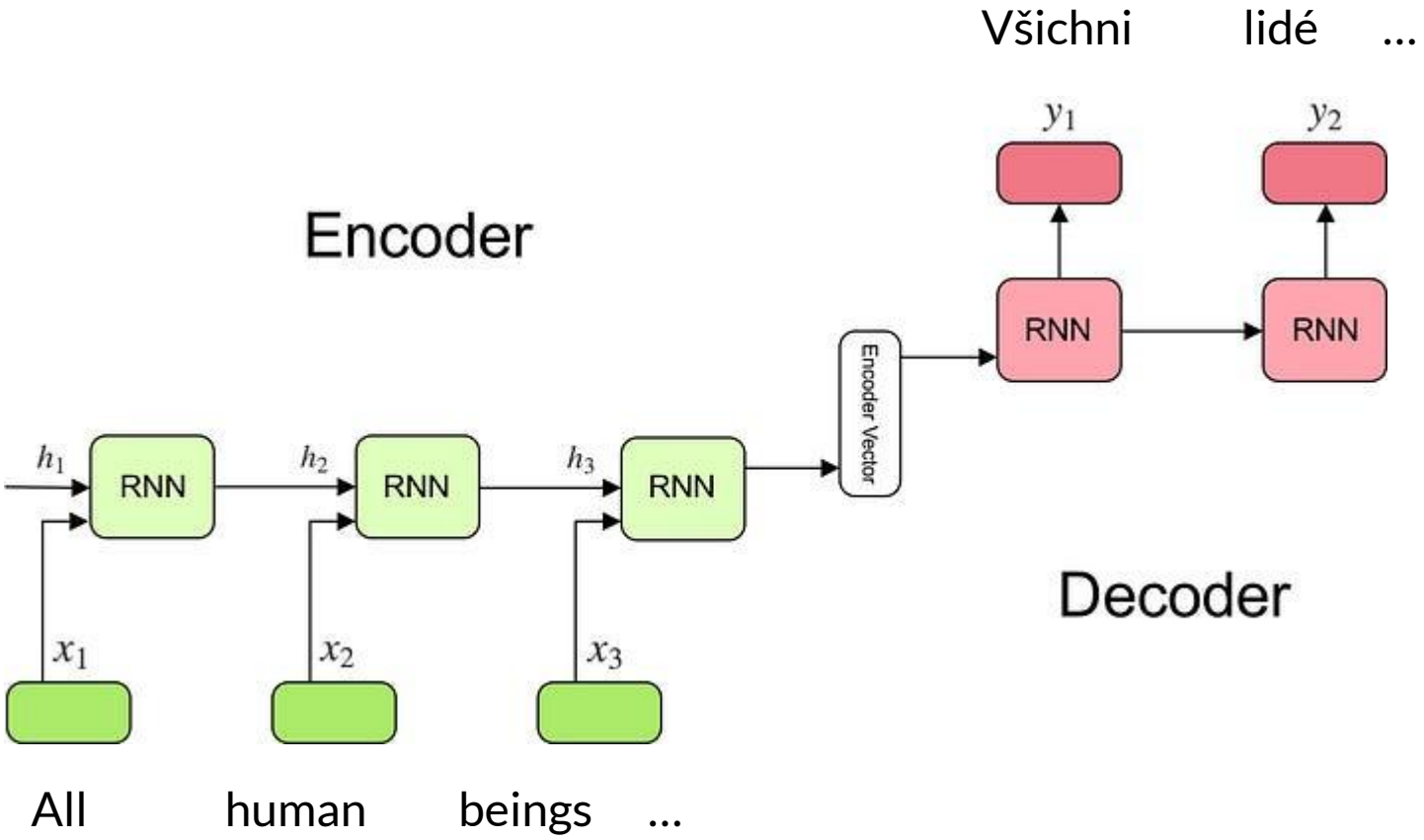
Siddharth M,
https://www.analyticsvidhya.com/blog/2021/07/feature-extraction-and-embeddings-in-nlp-a-beginners-guide-to-understand-natural-language-processing/

Recurrent Neural Network (RNN)



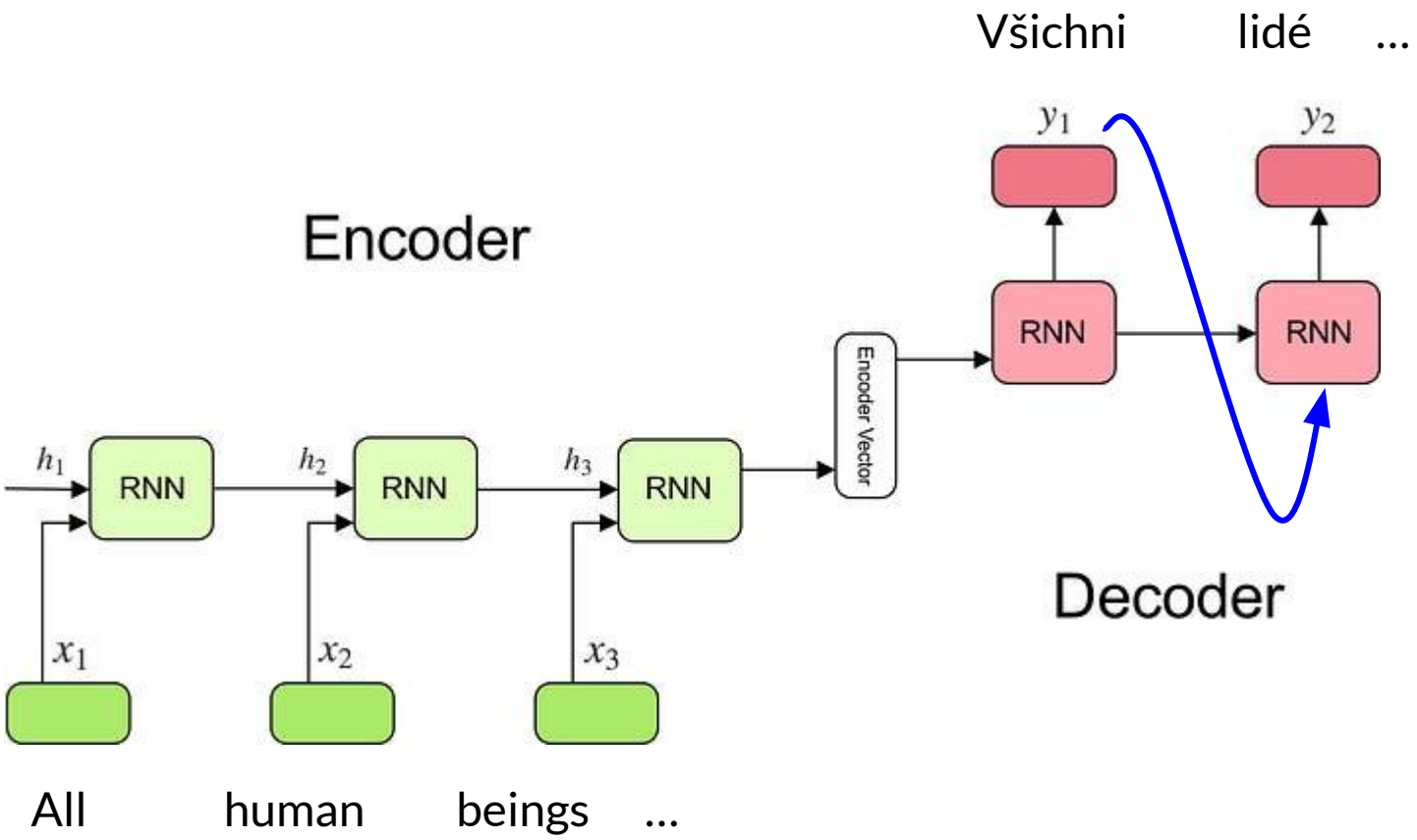
fdeloche, https://en.wikipedia.org/wiki/Recurrent_neural_network

Encoder-decoder (sequence-to-sequence)



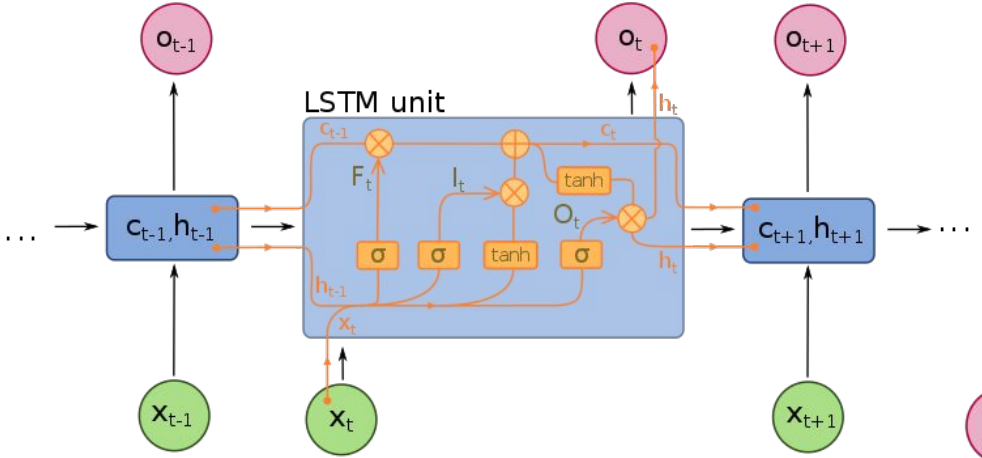
S. Kostadinov,
https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346

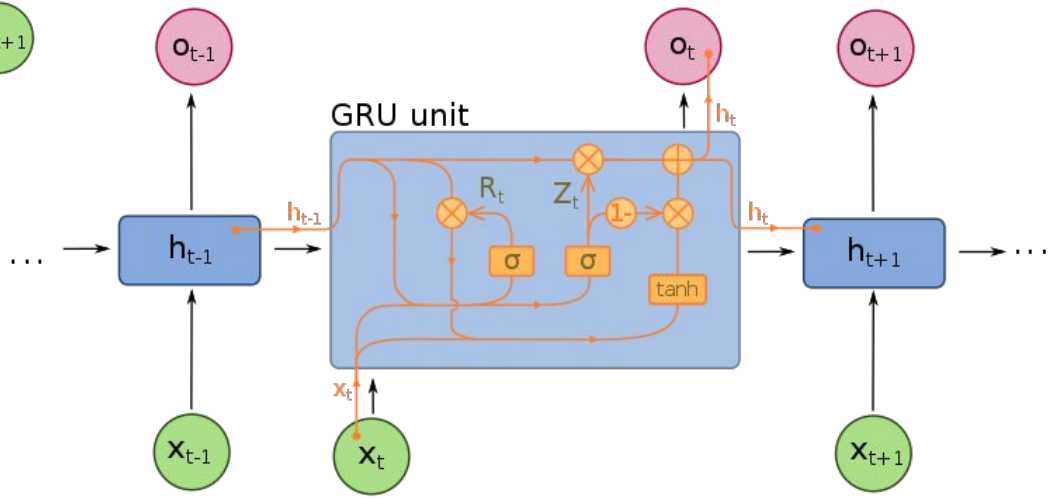
Based on previous output → Autoregressive Encoder-decoder



S. Kostadinov,
https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346

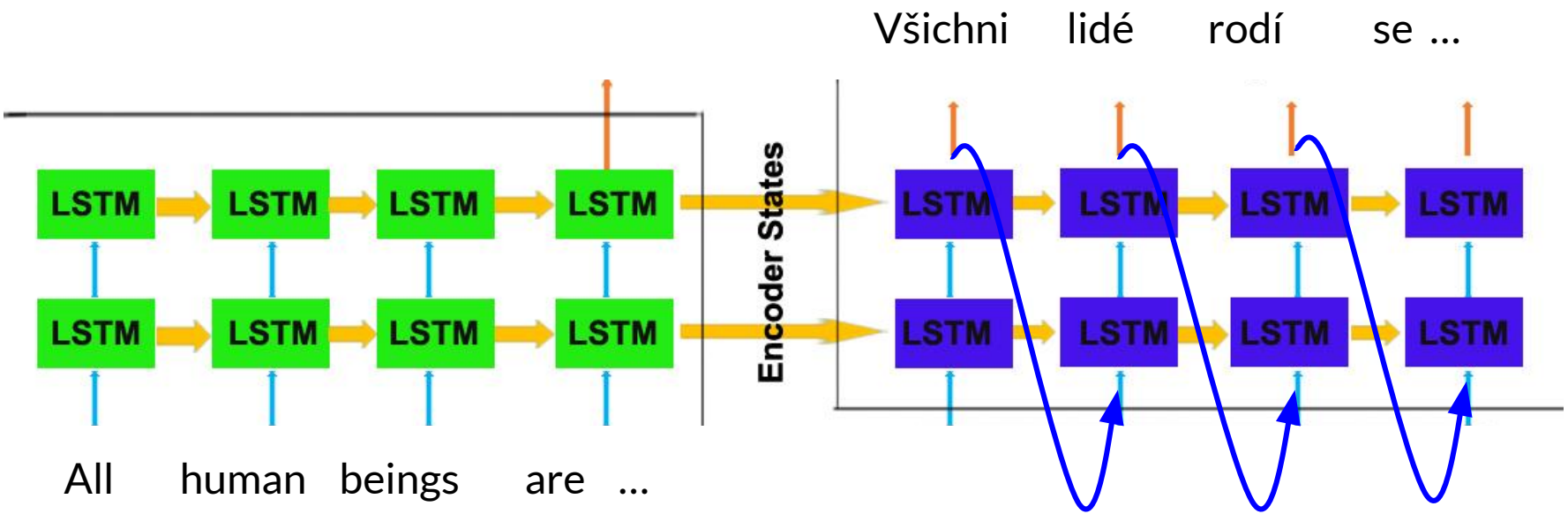
Vanishing/Exploding Gradient → RNNs with memory: LSTM, GRU





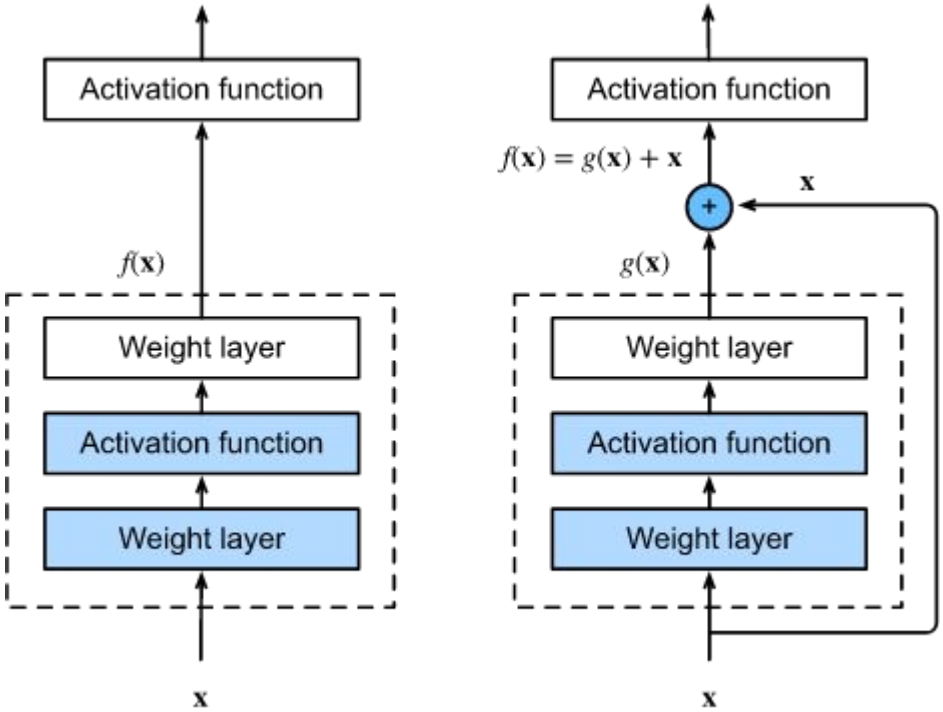
fdeloche, https://en.wikipedia.org/wiki/Recurrent_neural_network

Encode larger context → Stacked RNN



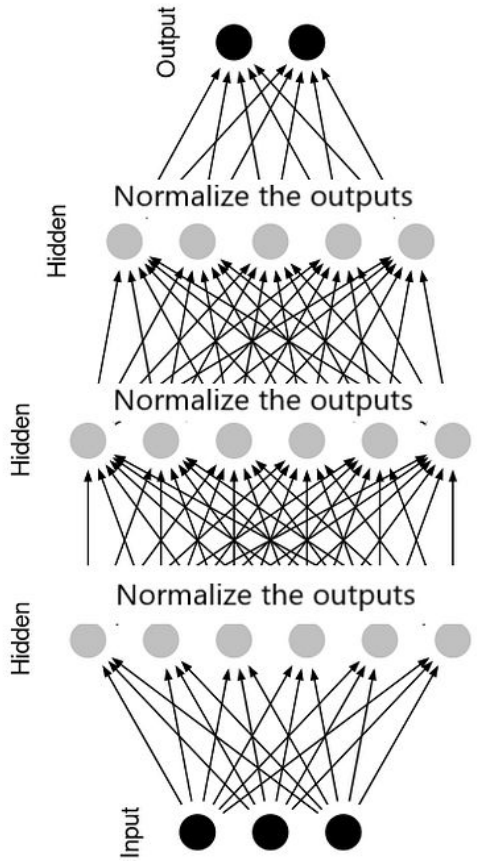
S. Jagadeesh,
https://www.analyticsvidhya.com/blog/2020/10/multivariate-multi-step-time-series-forecasting-using-stacked-lstm-sequence-to-sequence-autoencoder-in-tensorflow-2-0-keras/

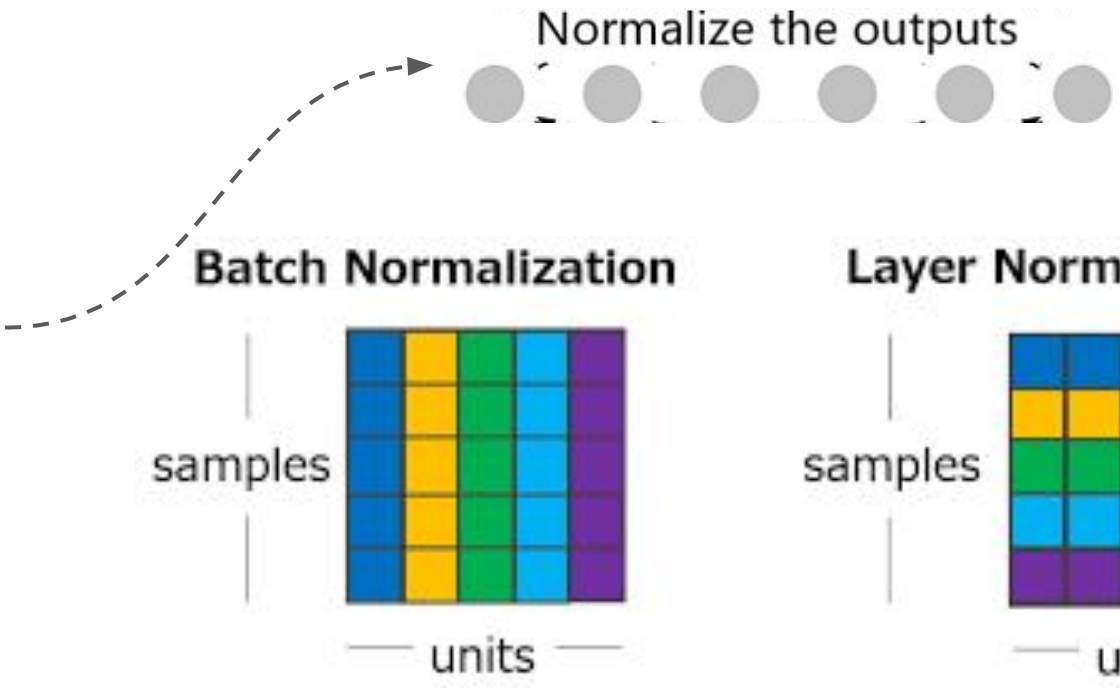
Vanishing/Exploding gradient → Residual Connections



A. Zhang+, https://d2l.ai/chapter_convolutional-modern/resnet.html

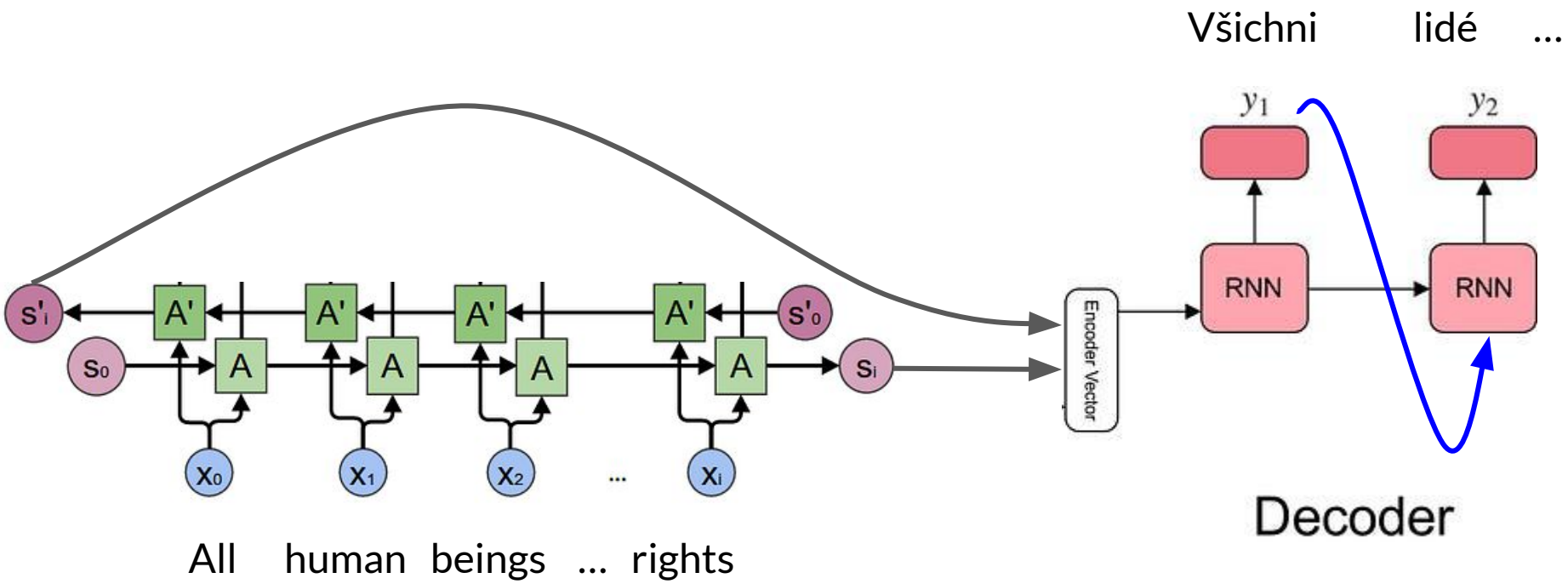
Covariate Shift → Batch / Layer Normalization





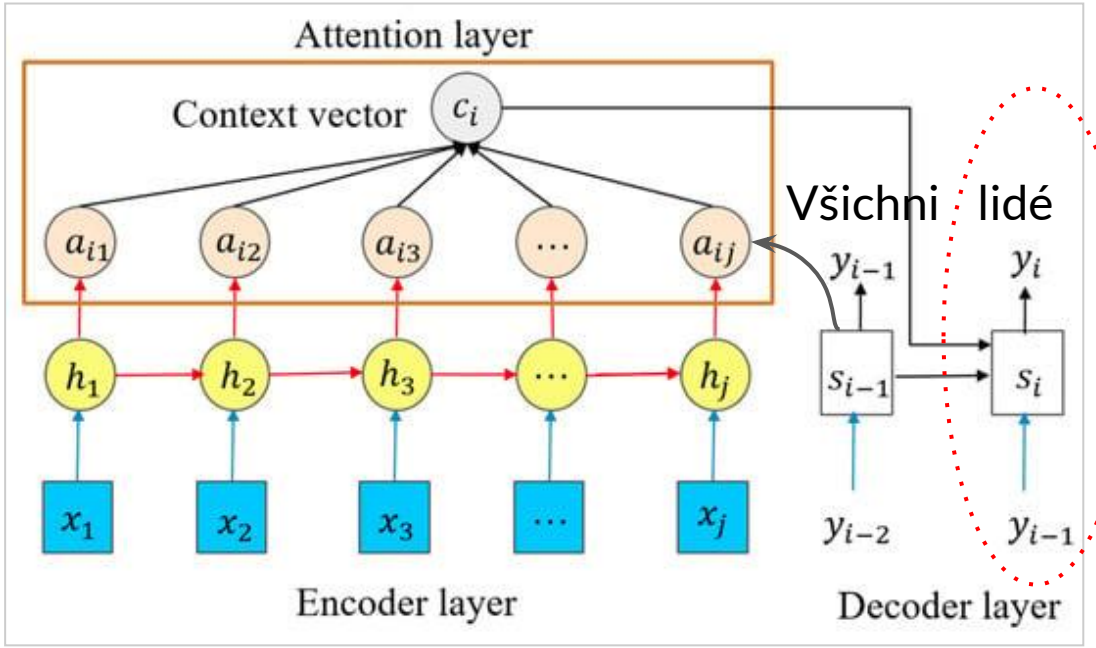
I. Rajagopal,
https://medium.com/@ilango100/batch-normalization-speed-up-neural-network-training-245e39a62f85
https://ai-pool.com/a/s/normalization-in-deep-learning

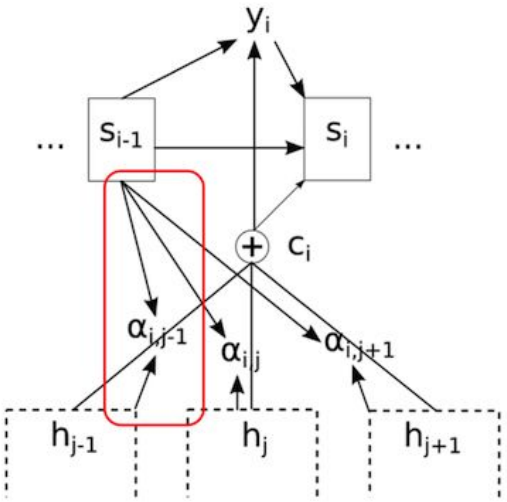
Encoding mostly the end → Bi-directional RNN



C. Lee,
https://towardsdatascience.com/understanding-bidirectional-rnn-in-pytorch-5bd25a5dd66

Selectively look at previous words' hidden states → Attention



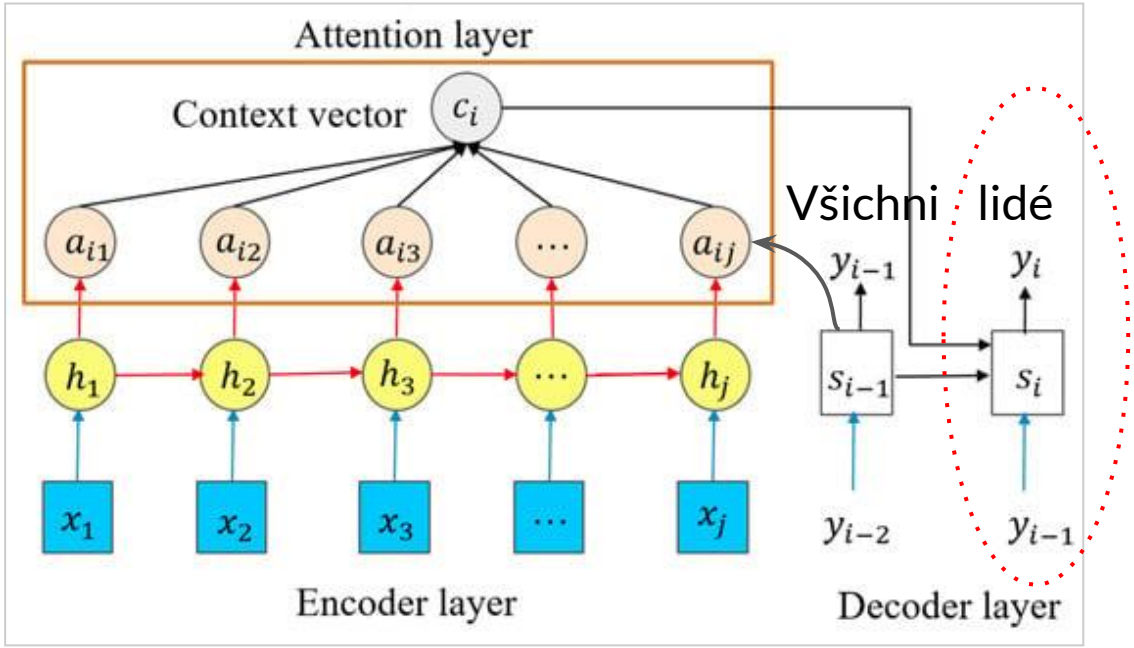


All human beings ... rights

$$a_{i,j} = \text{softmax}(e_{i,j})$$
$$e_{i,j} = v^T \tanh(W_1 h_i + W_2 s_{i-1})$$

D. Bahdanau+, https://humanbrain.gitbook.io/notes/notes/natural-language-processing/bahdanau_attention
M. Yang+, https://doi.org/10.3390/electronics10141657

Selectively look at previous words' hidden states → Attention

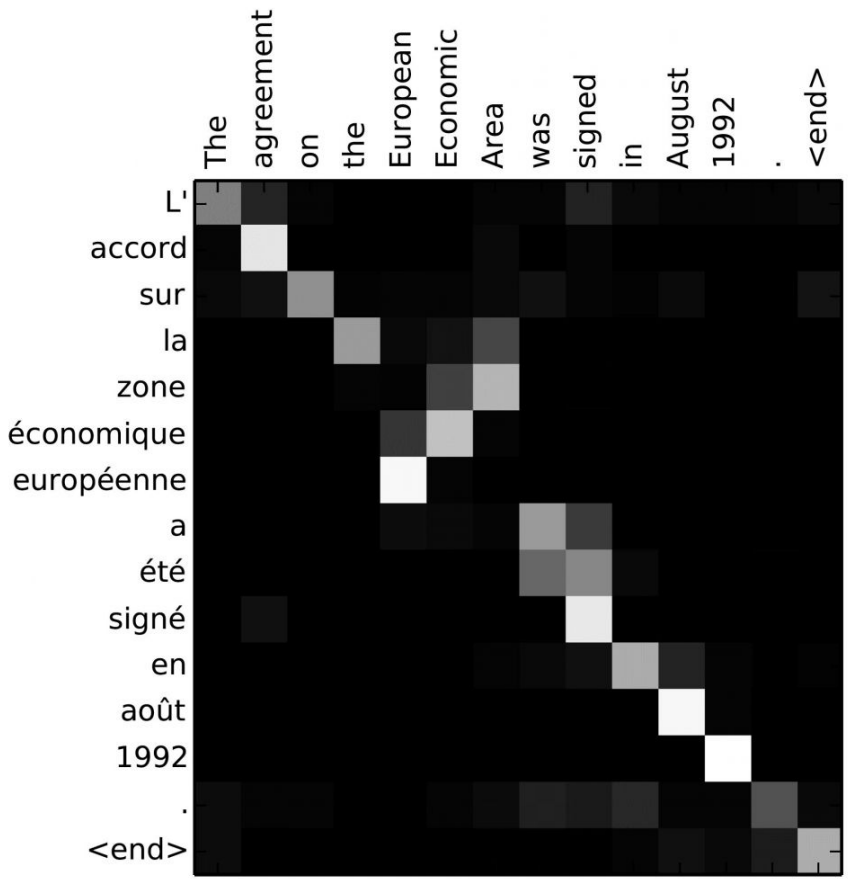


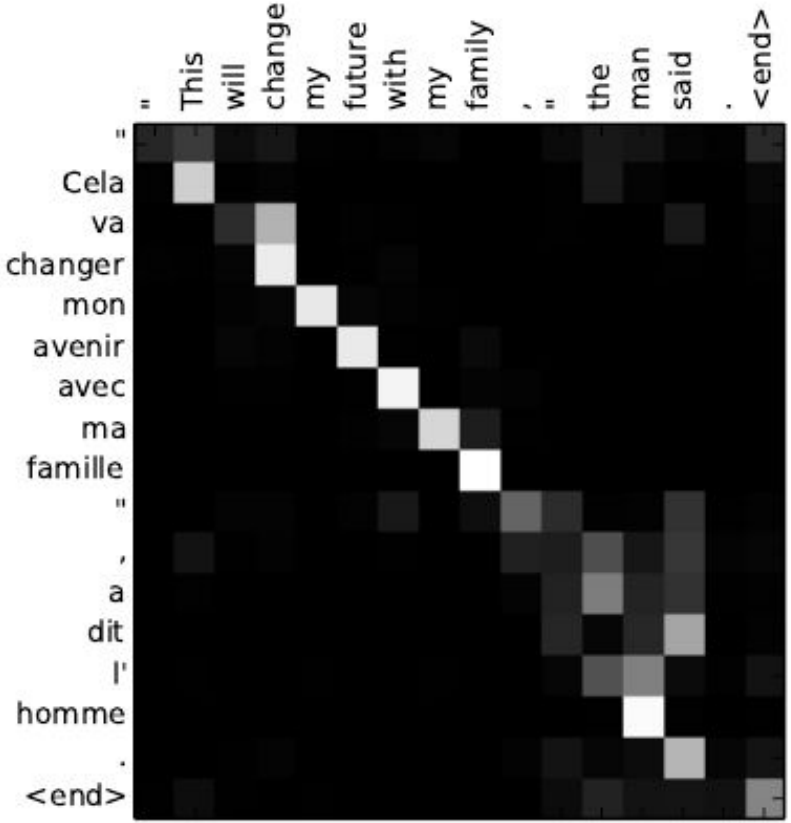
All human beings ... rights

- inspiration: word alignment
  - attention ~ soft alignment
- typically attends to word(s) being translated
  - "lidé" ~ "human beings"
- contextual word embs.
  - vs static word embeddings
  - hidden state ~ word representation in context
  - ELMo
- encoder final state useless
  - decode from a static state
  - use biRNN for hidden states

D. Bahdanau+, https://humanbrain.gitbook.io/notes/notes/natural-language-processing/bahdanau_attention
M. Yang+, https://doi.org/10.3390/electronics10141657

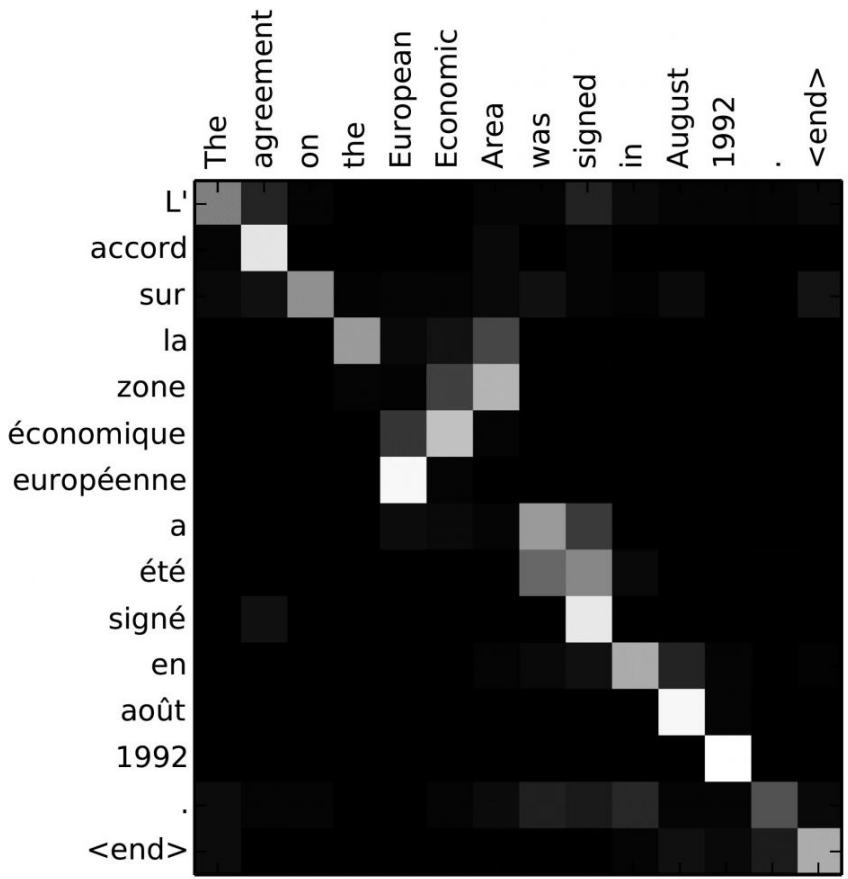
Visualisation of Attention





D. Bahdanau+, https://arxiv.org/abs/1409.0473

Visualisation of Attention



- this is single-head attention
  - typically attends to the word that is being translated
  - theoretically soft distribution
  - but typically very peaked
- also possible: multihead attention
  - one head attends to translated word
  - one head attends to preceding word
  - one head attends to sentence start
  - one head attends to sentence end
  - one head attends to main verb
  - one head attends to sentence subject
  - …
  - learned (emergent), not enforced!

D. Bahdanau+, https://arxiv.org/abs/1409.0473

A range of useful components and concepts

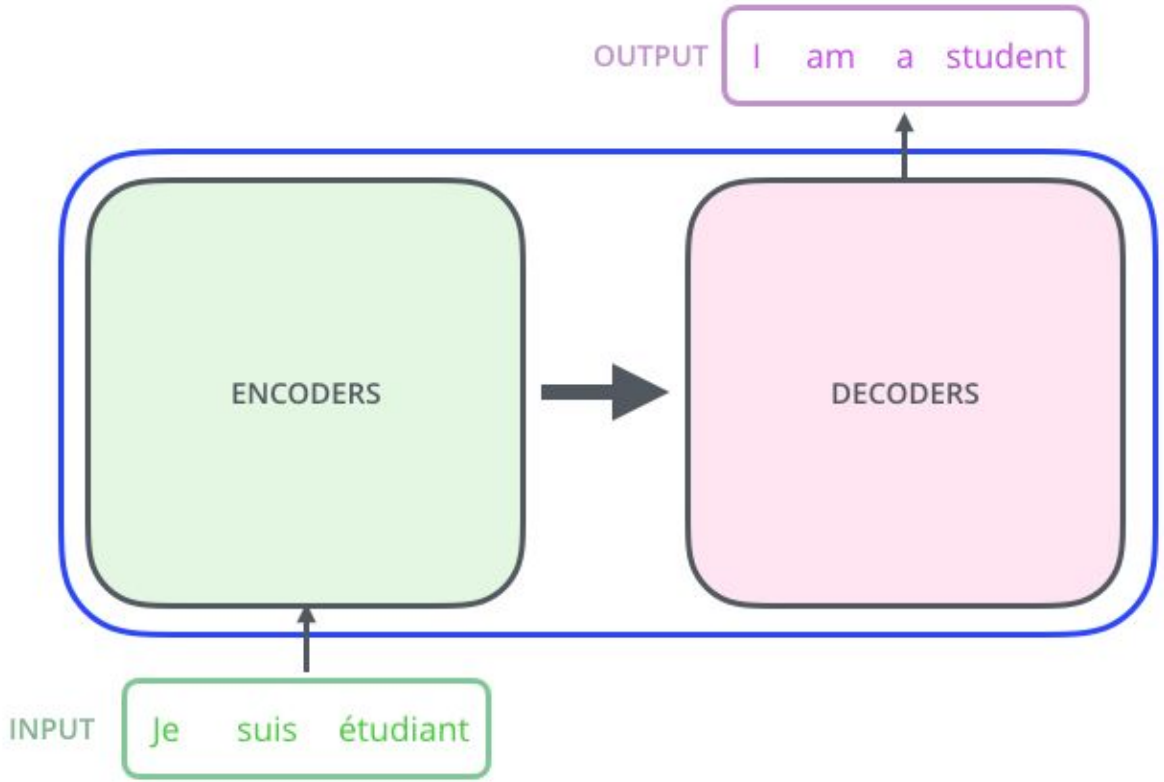
- NN cells (RNN, LSTM, GRU)
  - Note: CNNs possible but not common; sequence-to-label vs sequence-to-sequence
  - Note: Transformer is yet another alternative (to RNN or CNN) for sequence processing
- End-to-end
- Sequence-to-sequence
- Encoder-decoder
- Autoregressivness
- Stacking
- Residual Connections (Highway Networks)
- Batch/Layer Normalization
- Directionality
- Attention
  - very powerful → key component of Transformer (“Attention is All You Need”)

The Transformer

Transformer: end2end seq2seq

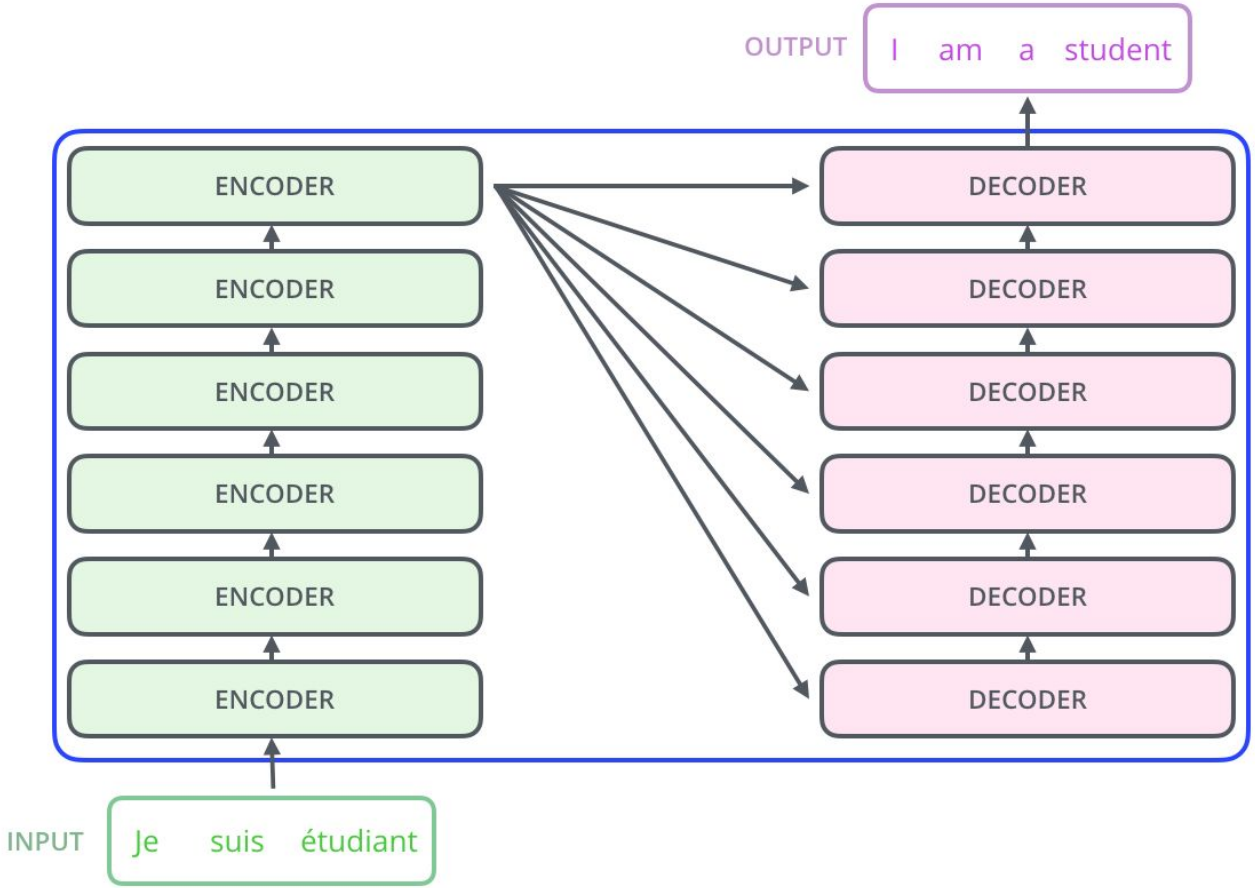
J. Alammar, https://jalammar.github.io/illustrated-transformer/

Transformer: encoder-decoder



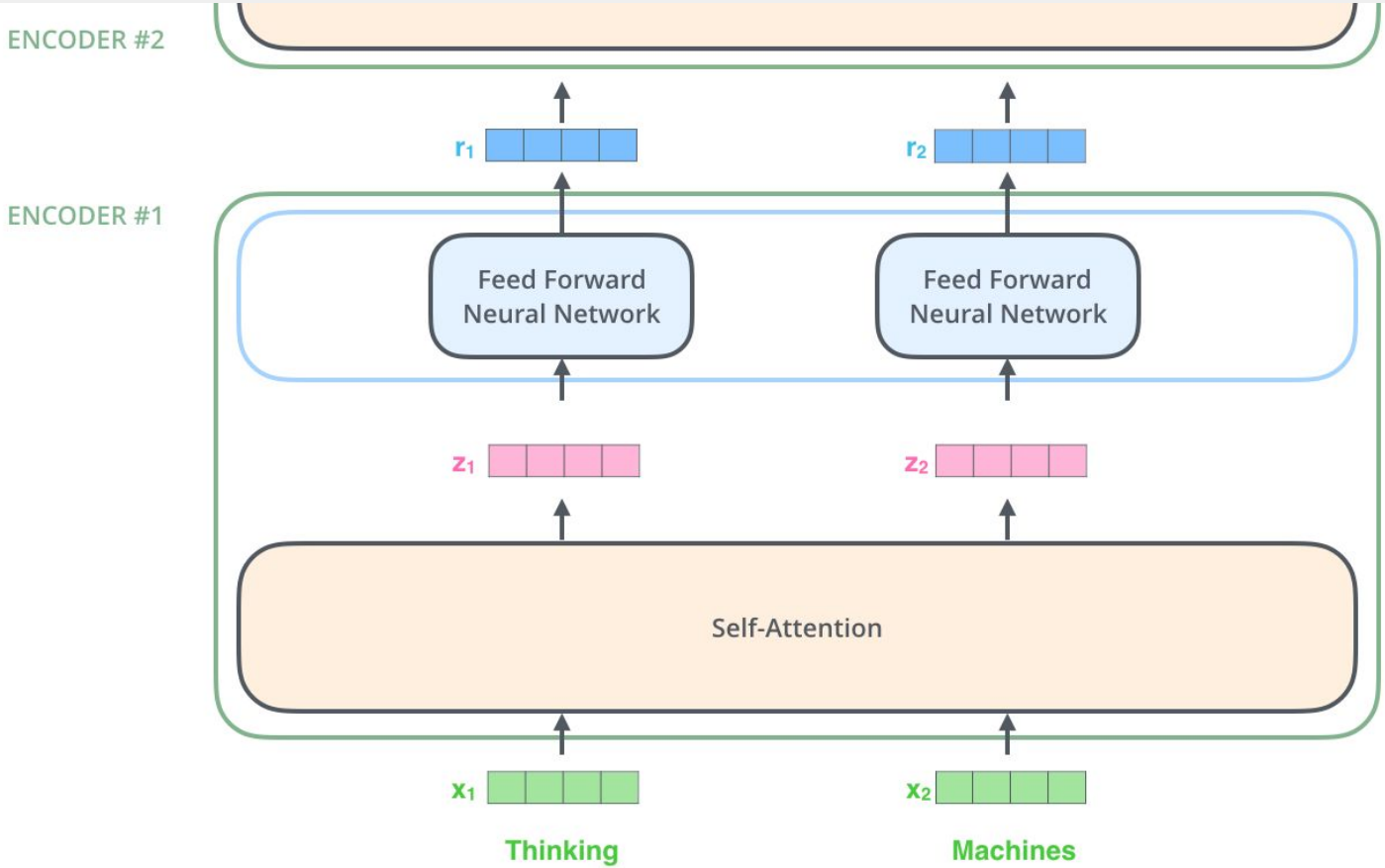
J. Alammar, https://jalammar.github.io/illustrated-transformer/

Transformer: stacking



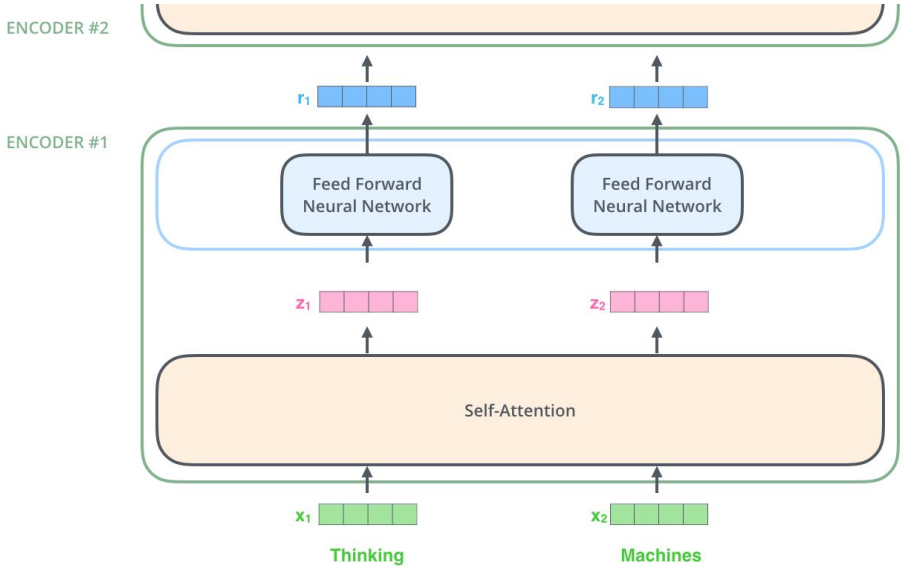
J. Alammar, https://jalammar.github.io/illustrated-transformer/

Transformer: encoder; hidden state ~ contextual word repres.



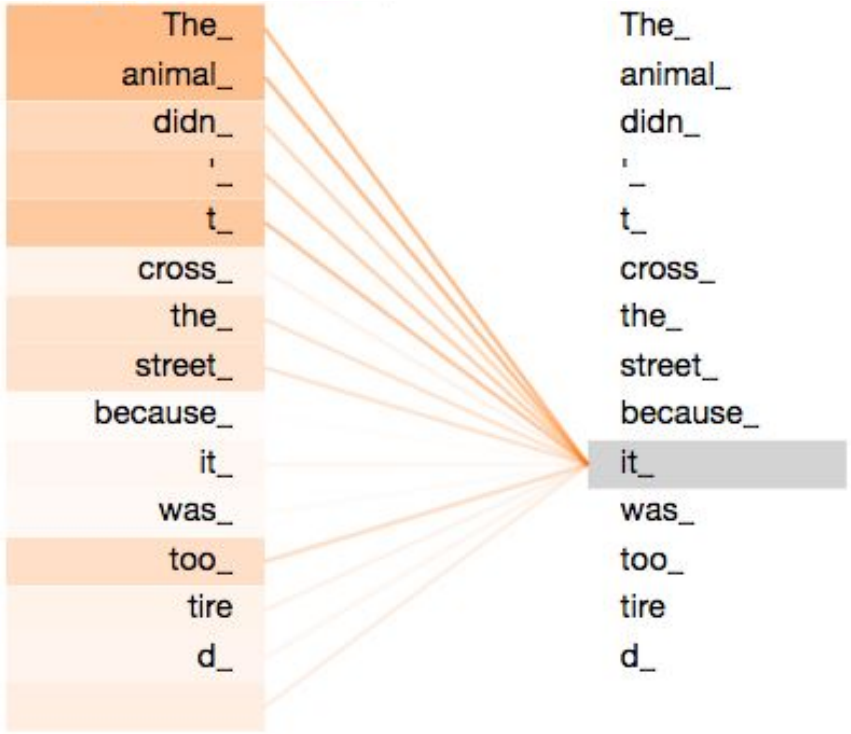
J. Alammar, https://jalammar.github.io/illustrated-transformer/

Transformer: self-attention (instead of RNN cell)



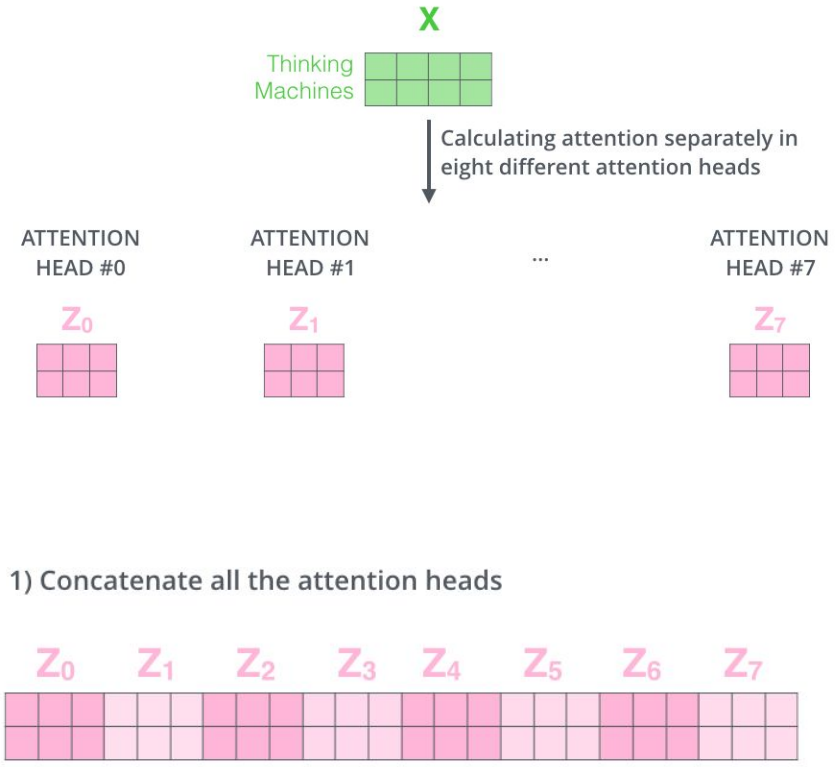
encoder input (x)

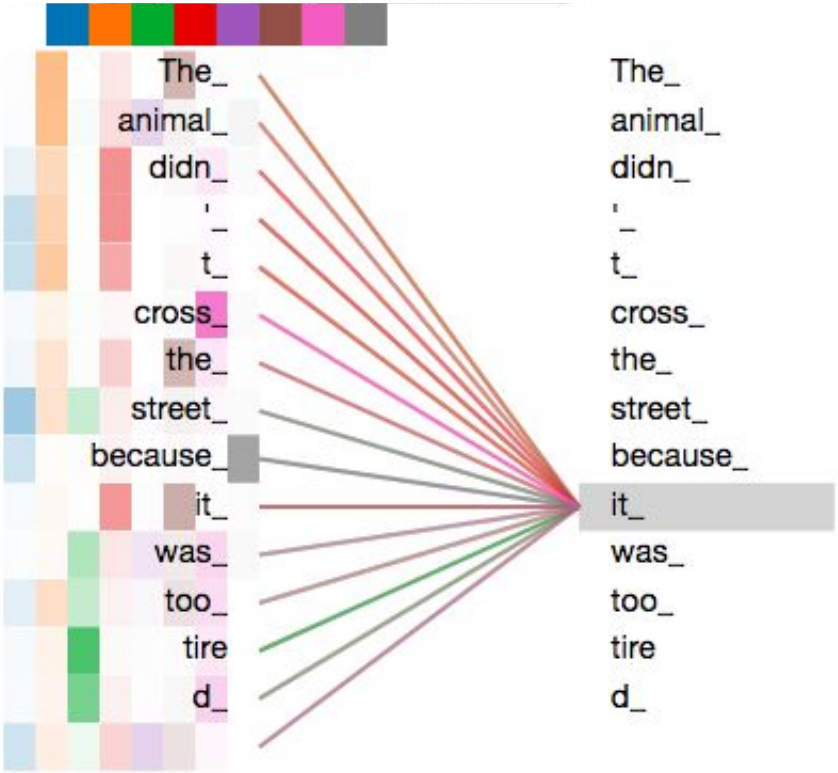
enc. internal state (z)



J. Alammar, https://jalammar.github.io/illustrated-transformer/

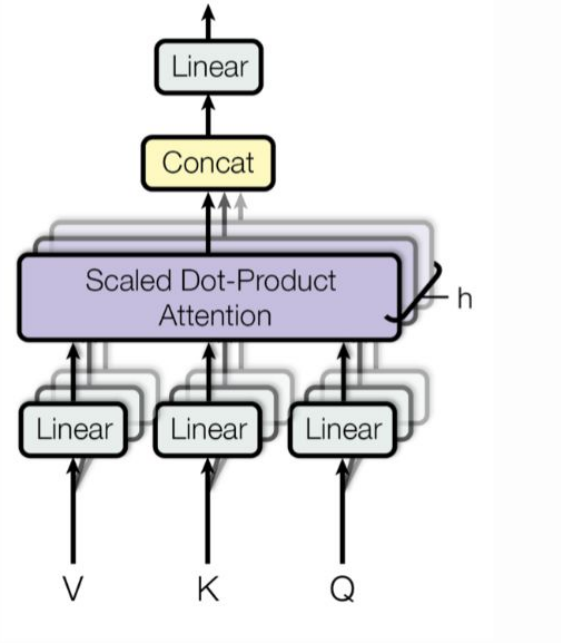
Transformer: multihead self-attention





J. Alammar, https://jalammar.github.io/illustrated-transformer/

Transformer: dot-product attention



Thinking

Input | Thinking | Machines

Embedding: X_1 X_2

Queries: q_1 q_2

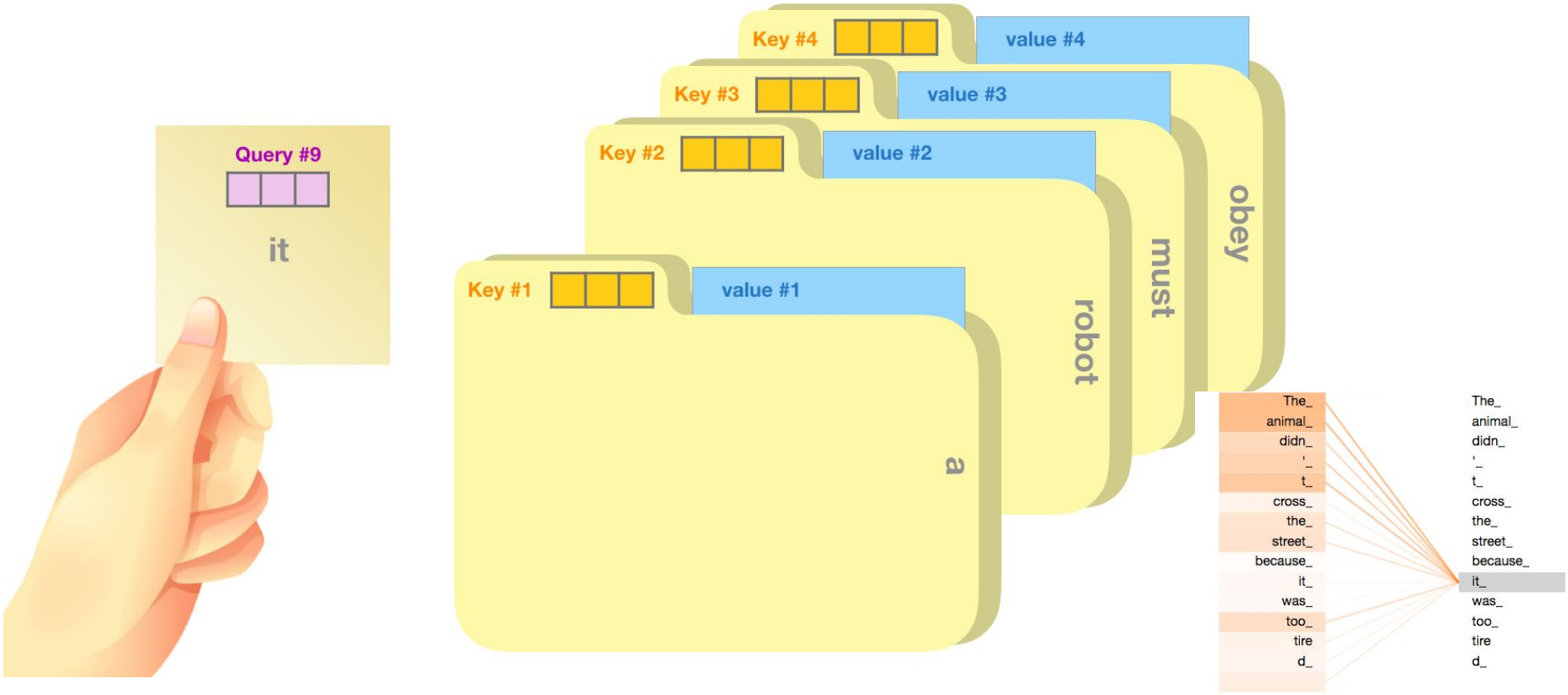
Keys: k_1 k_2

Values: v_1 v_2

W^Q W^K W^V

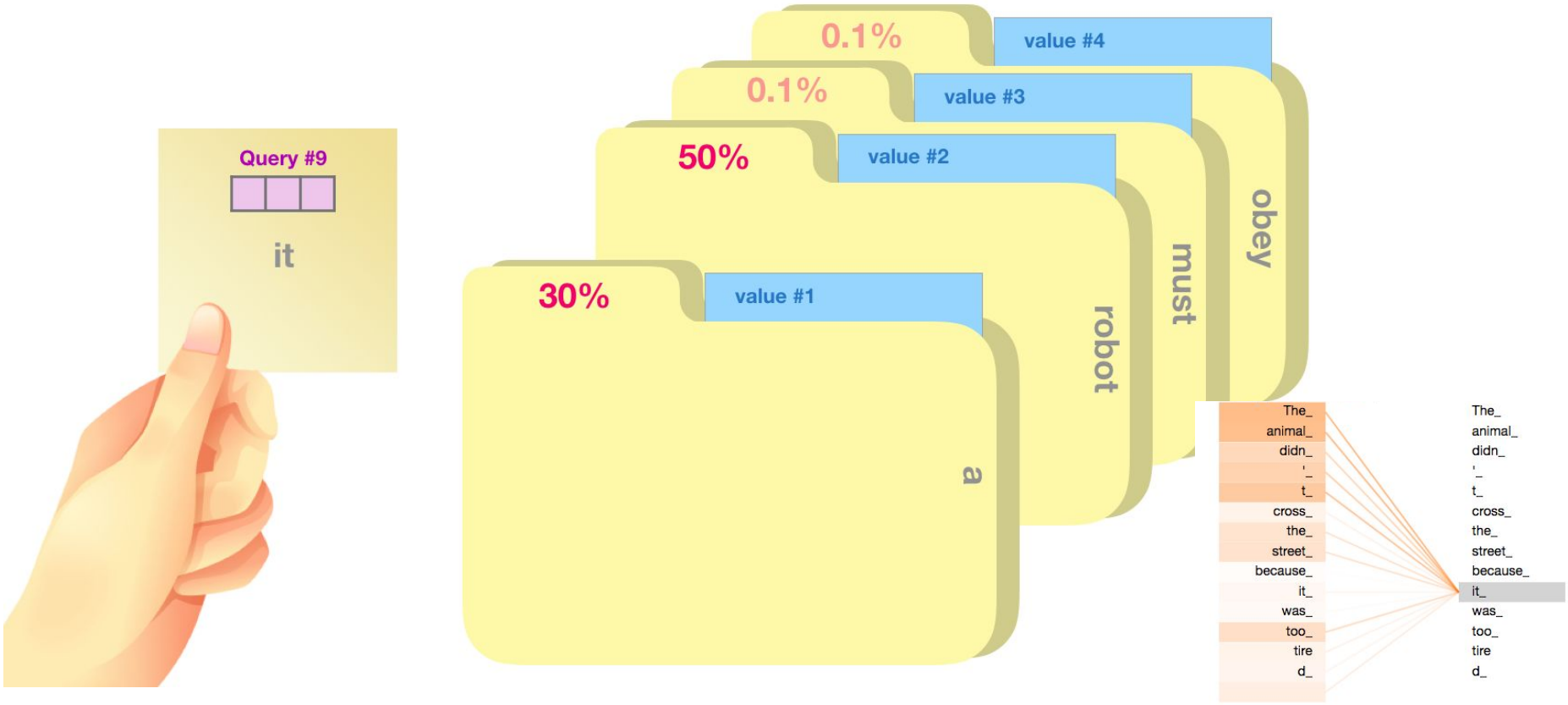
J. Alammar, https://jalammar.github.io/illustrated-gpt2

Transformer: dot-product attention



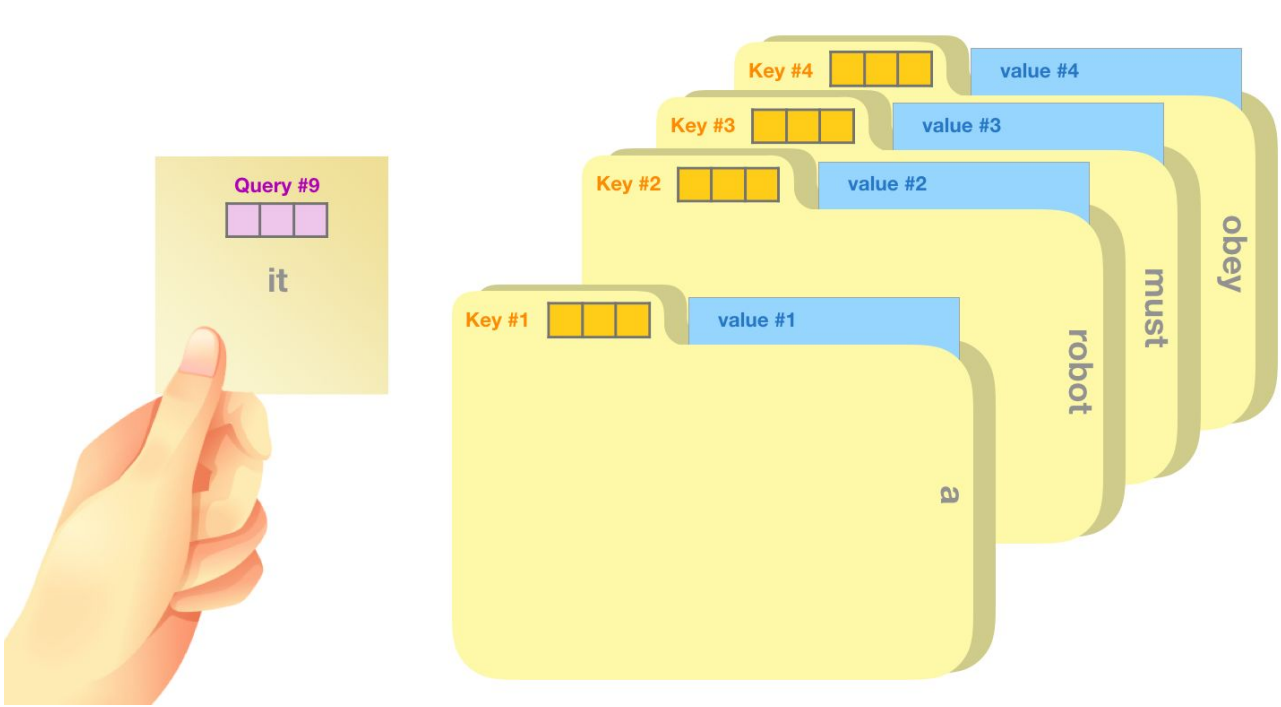
J. Alammar, https://jalammar.github.io/illustrated-gpt2

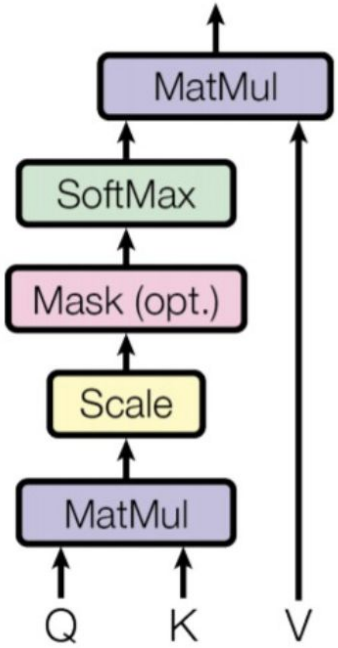
Transformer: dot-product attention



J. Alammar, https://jalammar.github.io/illustrated-gpt2

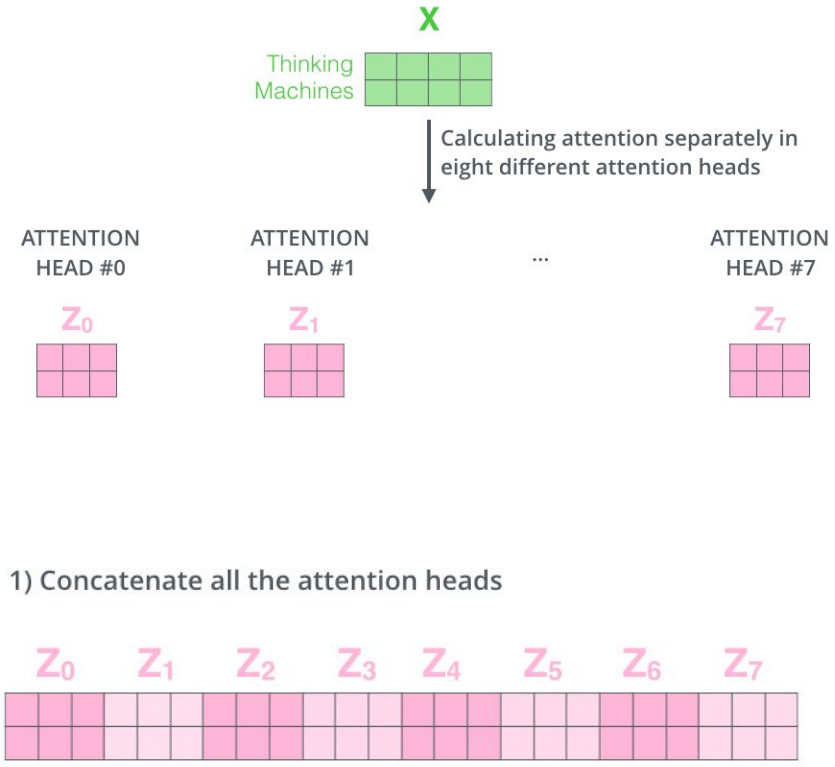
Transformer

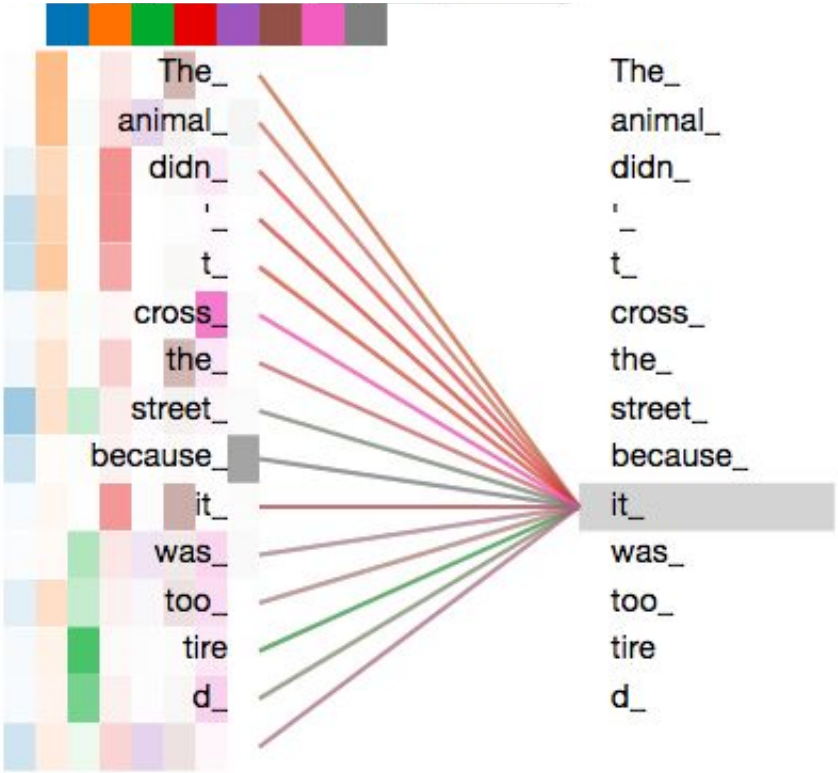




J. Alammar, https://jalammar.github.io/illustrated-transformer/

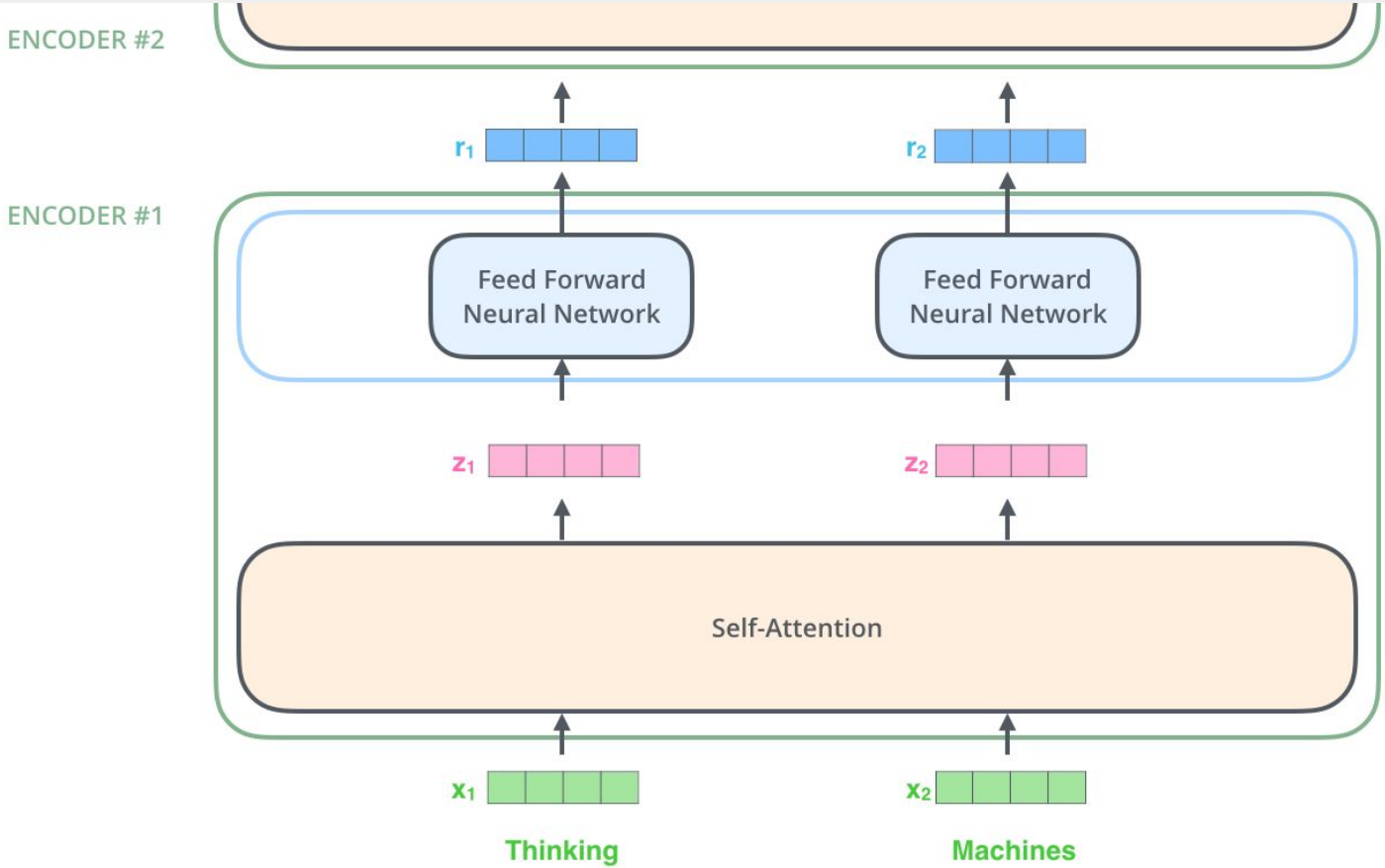
Transformer: multihead self-attention





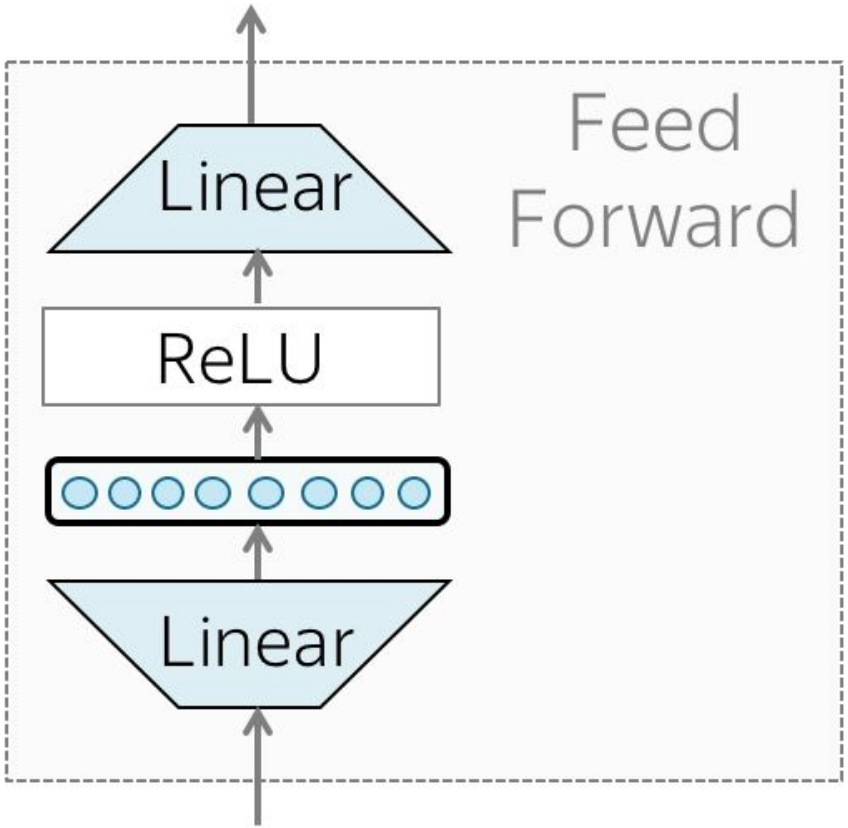
J. Alammar, https://jalammar.github.io/illustrated-transformer/

Transformer: encoder; hidden state ~ contextual word repres.



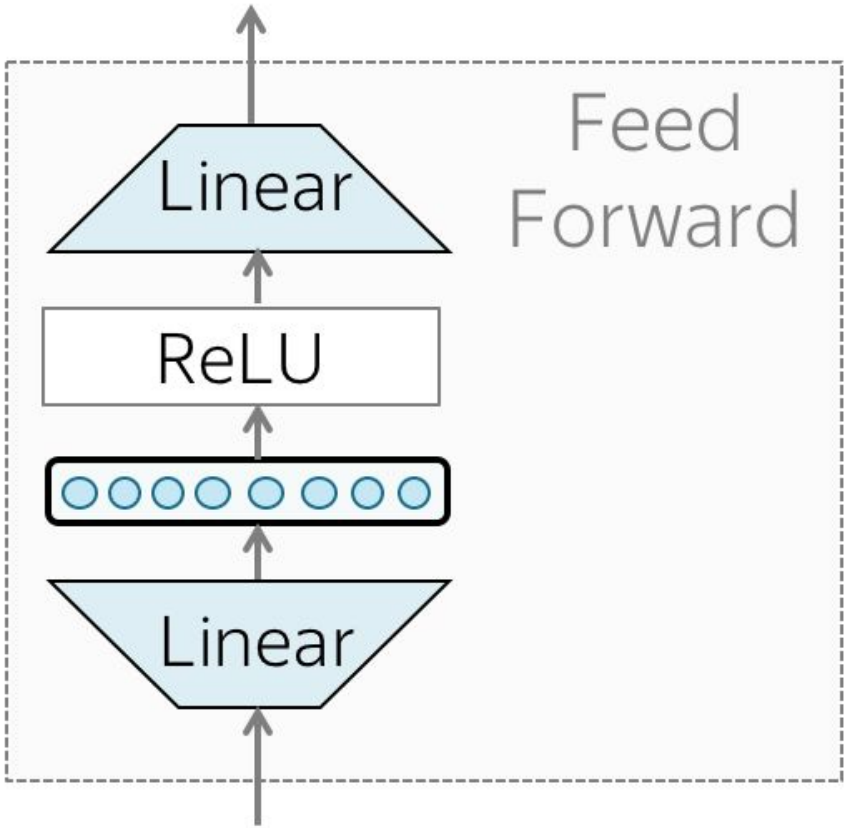
J. Alammar, https://jalammar.github.io/illustrated-transformer/

Transformer FFN: 1 hidden layer, “expand and contract” (4x)



L. Voita, https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html

Transformer FFN: 1 hidden layer, “expand and contract” (4x)



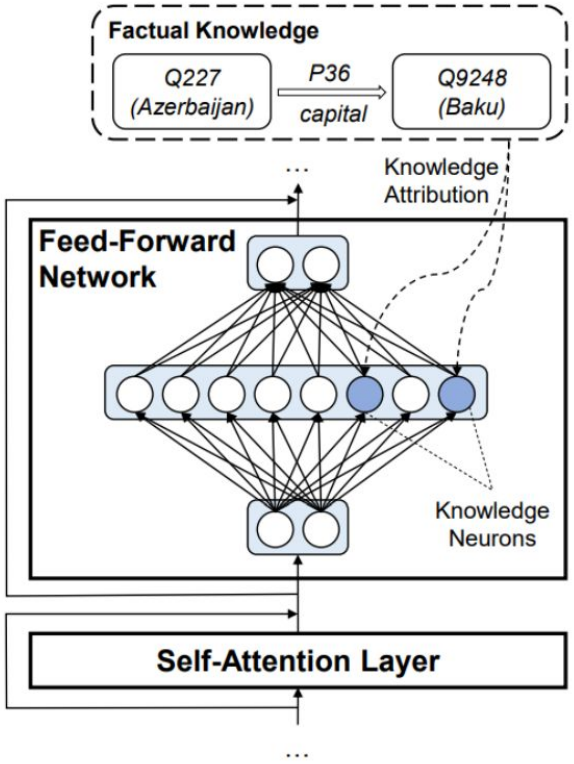
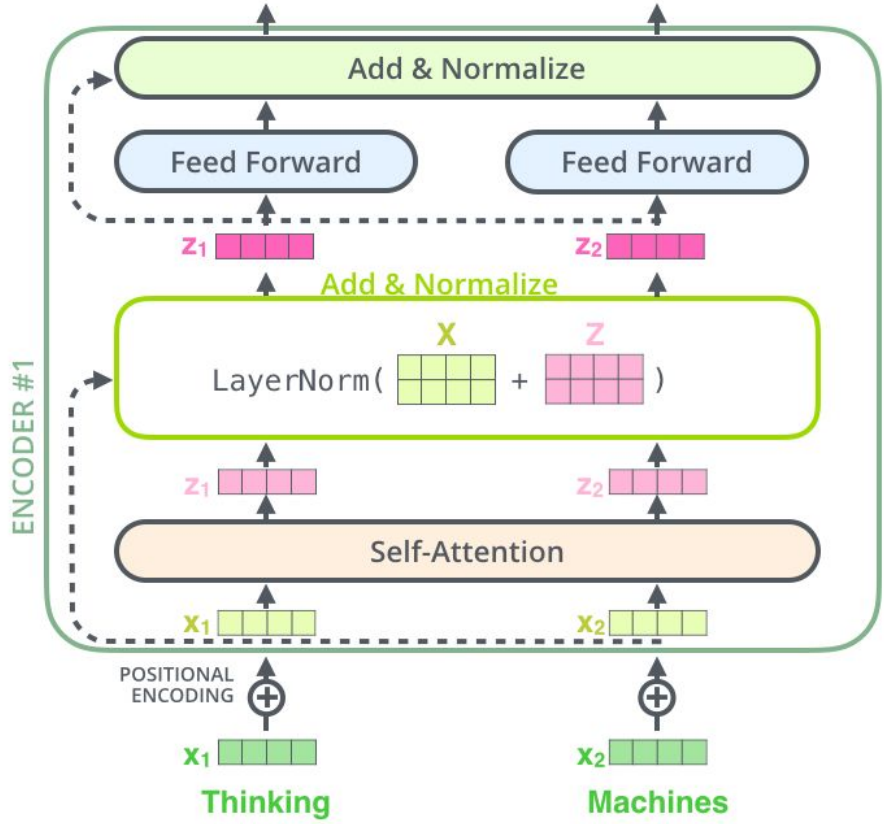


Figure 1: We aim to identify knowledge neurons correlated to a relational fact through knowledge attribution.

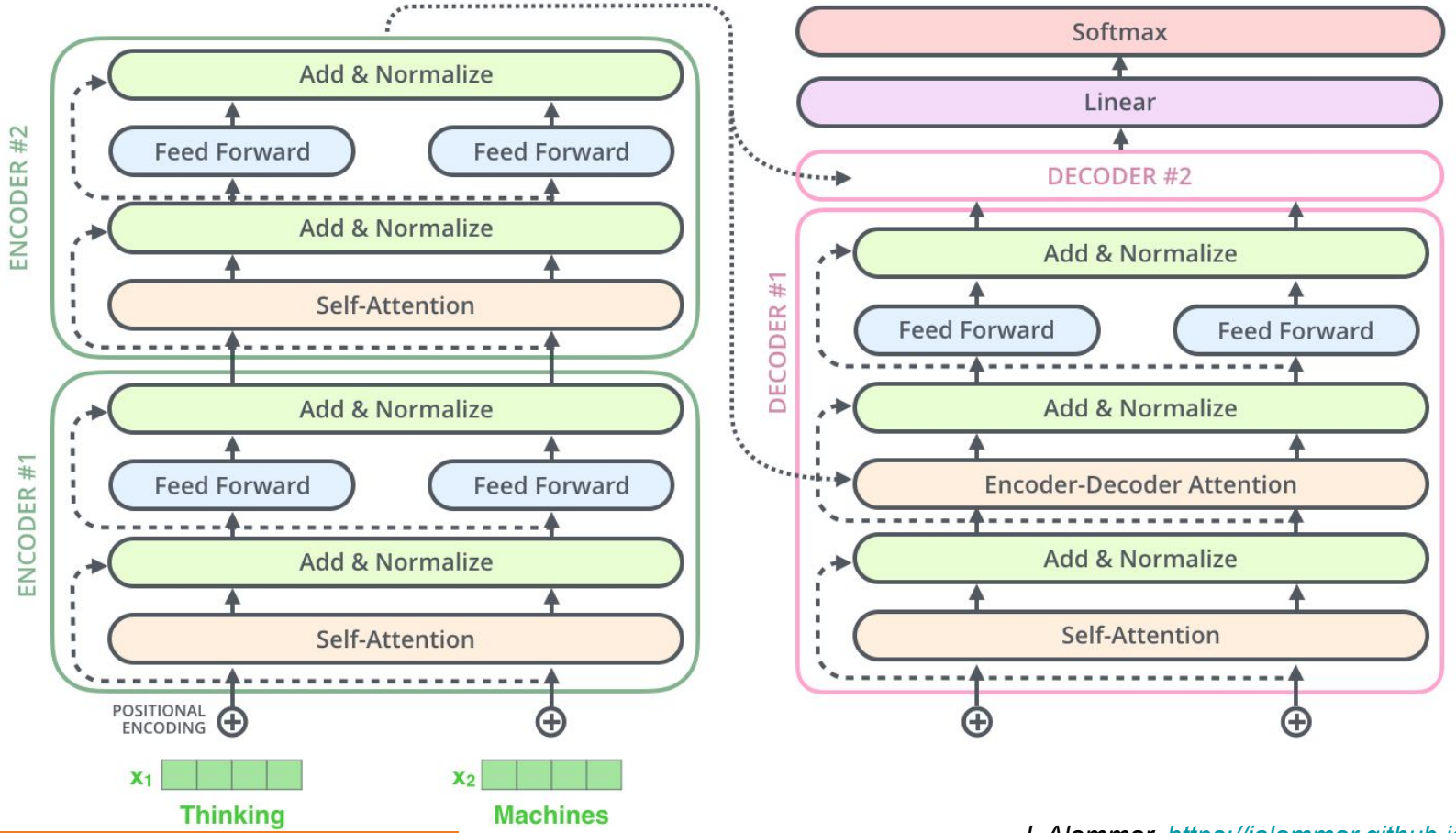
D. Dai+, https://aclanthology.org/2022.acl-long.581/

Transformer: full encoder block (residual connections, layer norm)



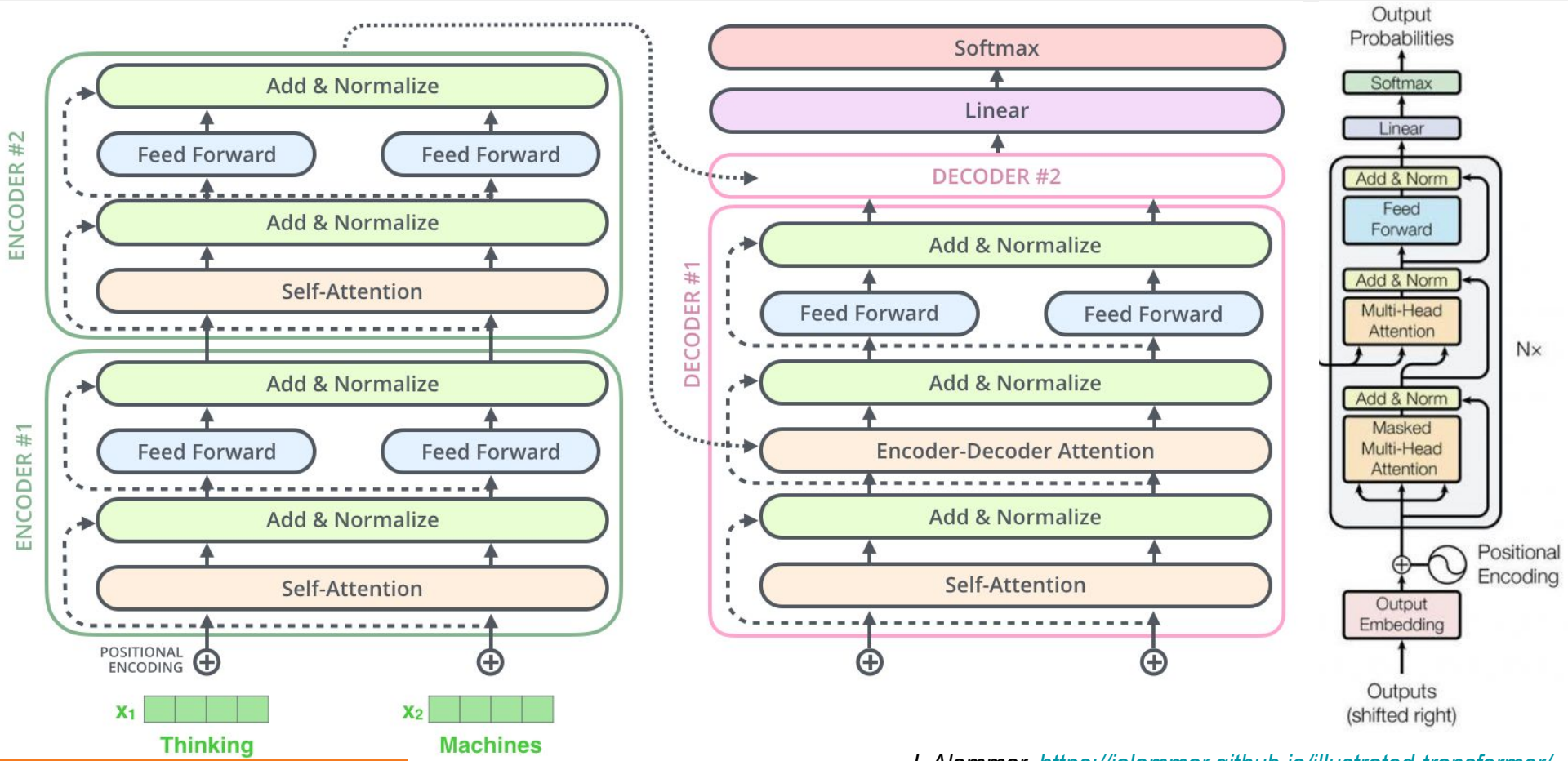
J. Alammar, https://jalammar.github.io/illustrated-transformer/

Transformer: encoder and decoder



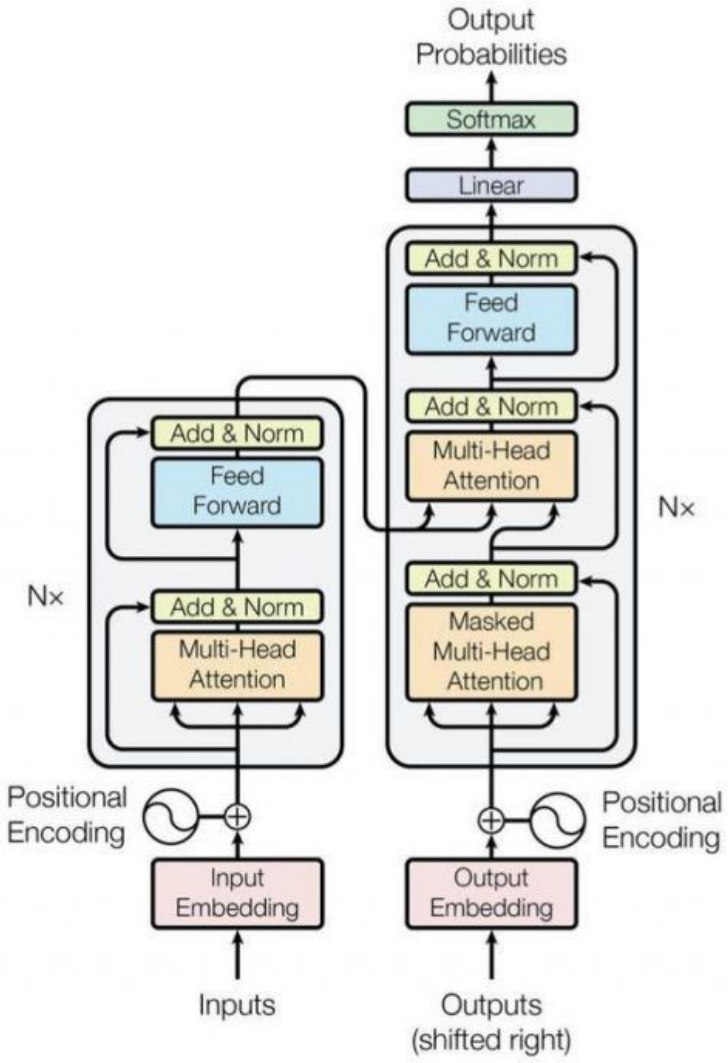
J. Alammar, https://jalammar.github.io/illustrated-transformer/

Transformer: encoder and decoder



J. Alammar, https://jalammar.github.io/illustrated-transformer/

Transformer



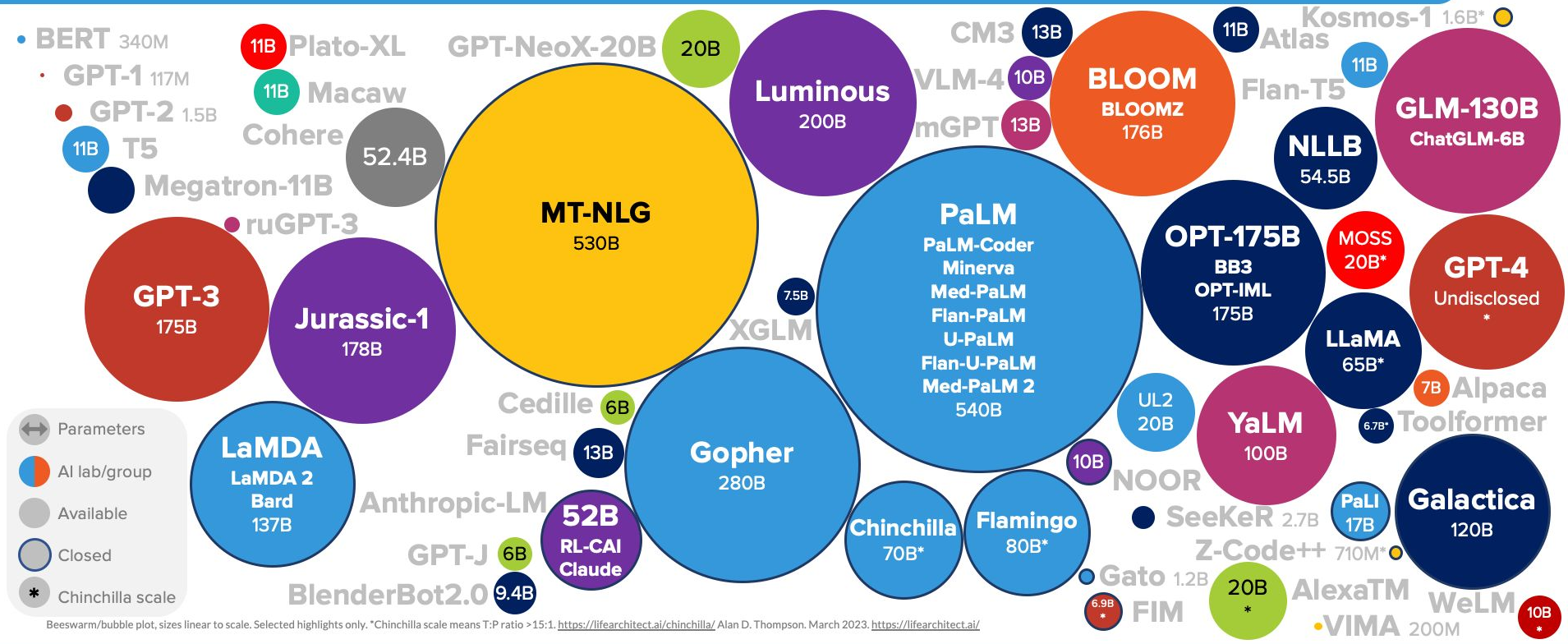
J. Alammar,
https://jalammar.github.io/illustrated-transformer/

Transformers Models and Applications

Applications

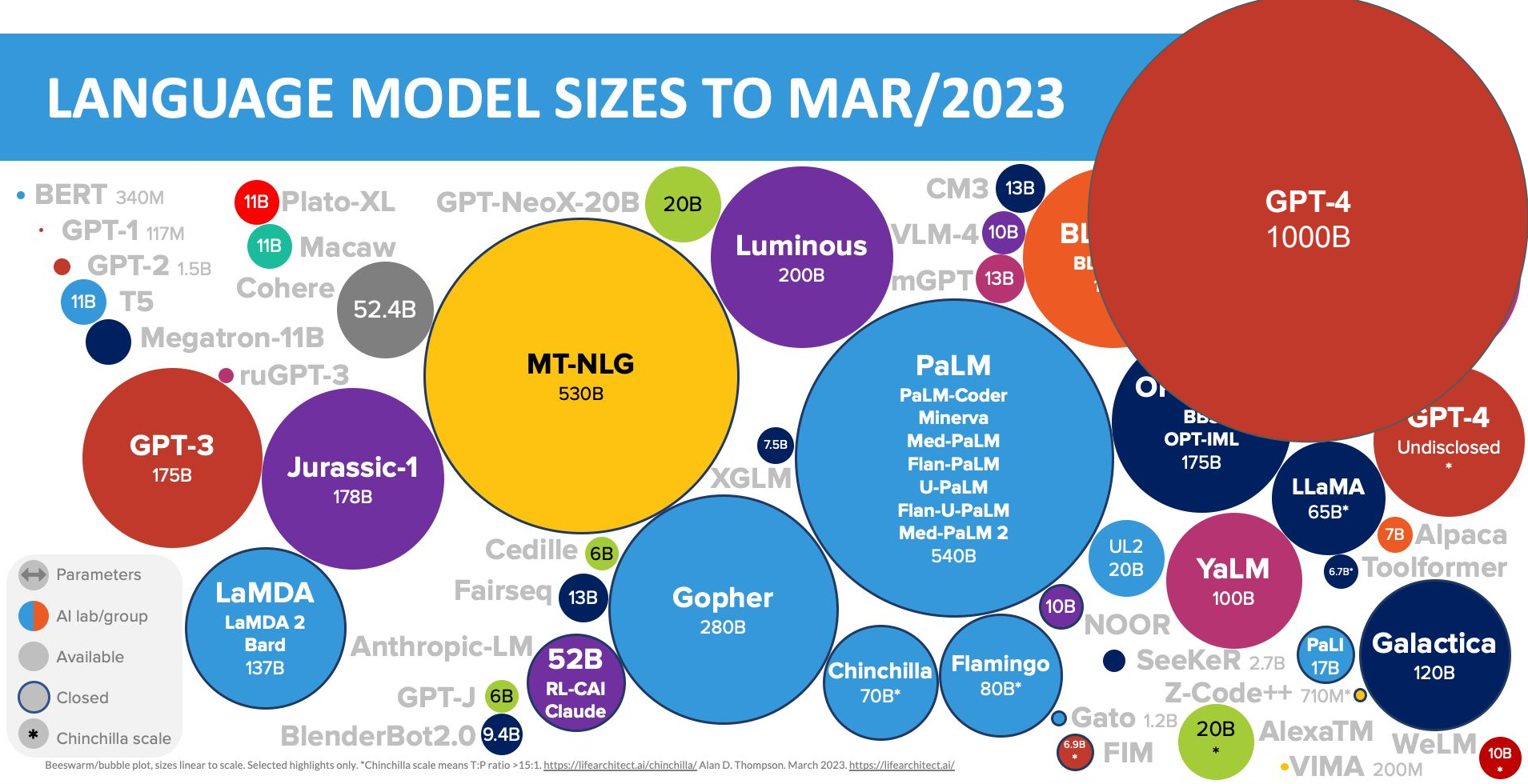
- Sequence-to-sequence models
  - Machine Translation
    - parallel data (sentence → sentence' / paragraph → paragraph')
  - Text Summarisation
    - text → summary
  - …
- Language Modelling
  - self-supervised (text → text)
  - next word prediction: *I woke up in the morning and went to the…*
    - autoregressive text generation
  - masked word prediction: *I woke up in the [MASK] and went to the bathroom.*
    - fill missing word, error correction
    - text representation (~ semantic tasks)
      - static word embeddings → contextual word embeddings → sentence embedding
      - text classification (part of speech, sentiment…)
- …

LANGUAGE MODEL SIZES TO MAR/2023



Beeswarm/bubble plot, sizes linear to scale. Selected highlights only. *Chinchilla scale means T:P ratio >15:1. https://lifearchitect.ai/chinchilla/ Alan D. Thompson. March 2023. https://lifearchitect.ai/

LANGUAGE MODEL SIZES TO MAR/2023



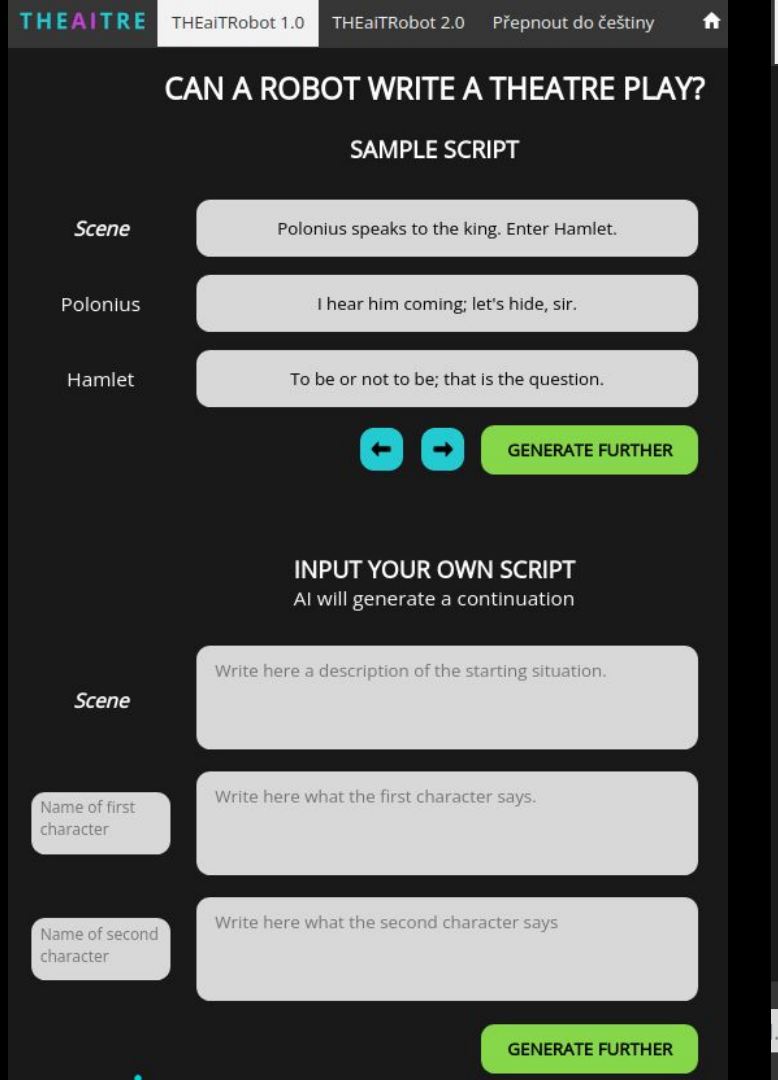
Beeswarm/bubble plot, sizes linear to scale. Selected highlights only. *Chinchilla scale means T:P ratio >15:1. https://lifearchitect.ai/chinchilla/ Alan D. Thompson. March 2023. https://lifearchitect.ai/

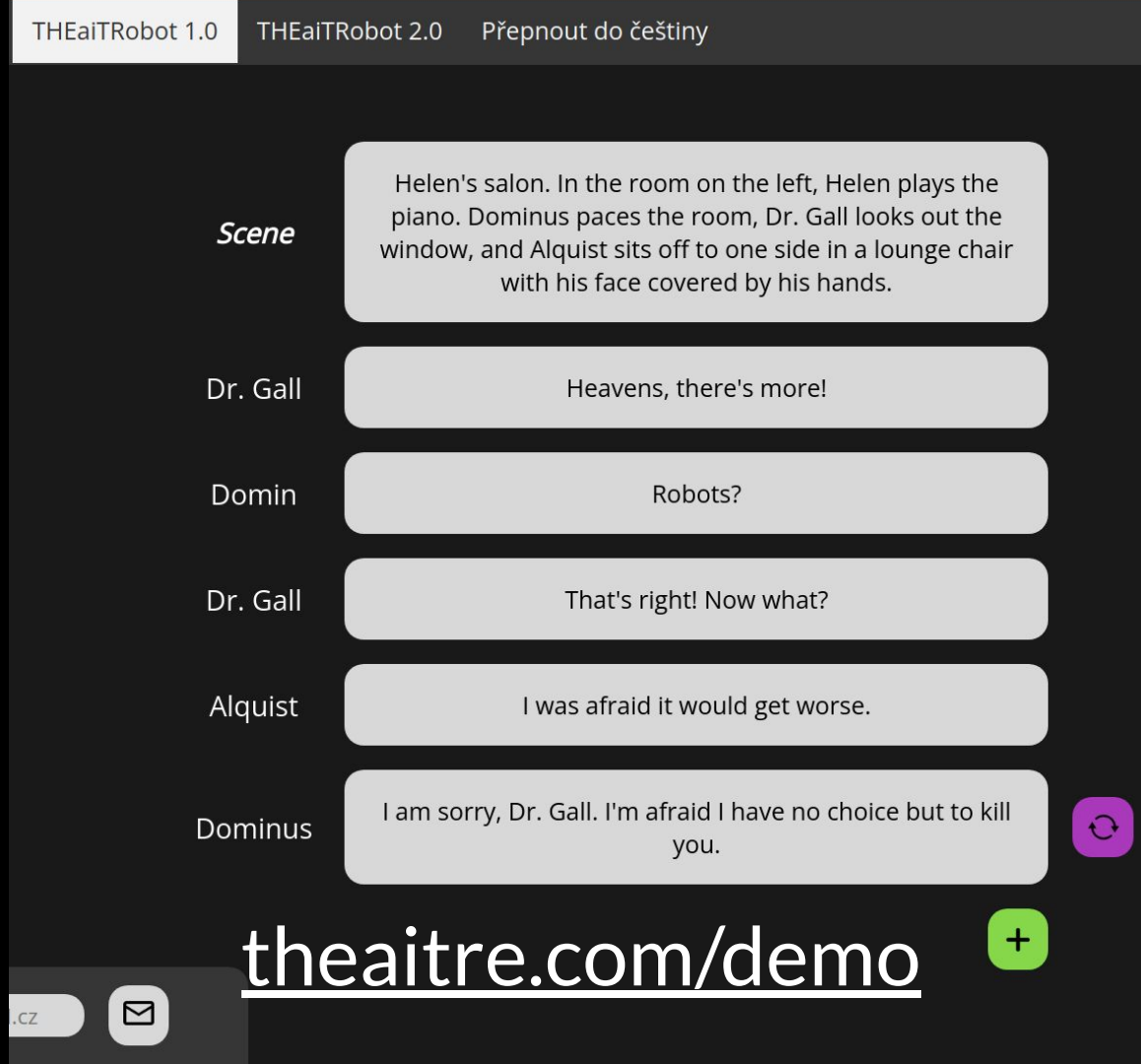
Notable language models

- masked
  - BERT, Multilingual BERT, RoBERTa, XLM-RoBERTa, BART
- generative
  - OpenAI: GPT models (GPT-1 … GPT-4, ChatGPT)
  - Meta: OPT, LLaMA → Alpaca, Vikuna, gpt4all
  - academia: HPLT, OpenGPT-X, OpenAssistant (all in progress)
- model distillation (e.g. DistillBERT)
  - teacher-student: train a large model, teach a small model to simulate the large model
- going beyond naturally occurring text (i.e. crawling the internet) in GPT-3.5
  - instructions (InstructGPT): “Summarize the following text: …”
  - source codes (Copilot): “Write a Python script to sort an array…”
  - reinforcement learning with human feedback (RLHF): 👍👎 “correct answer is: …”
- going beyond text → multimodal (images), external tools, embodiment (robots)
  - GPT-4, Retrieval-enriched LMs, Bing AI Bot, ToolFormer, ChatGPT plugins, PaLM-E

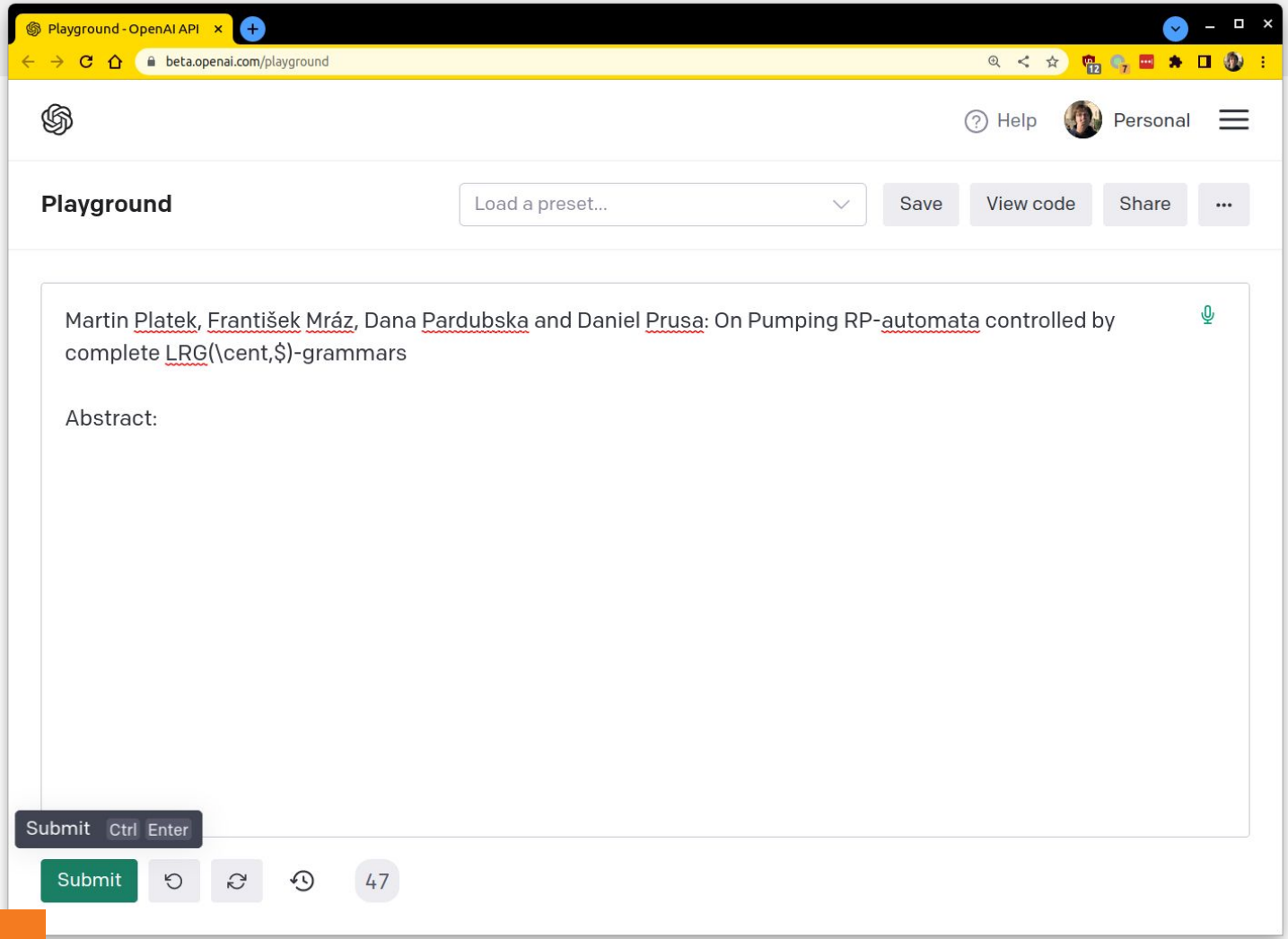
(Some) Text Generation Free Online Demos

- our demo
  - THEaiTRobot demo
    - GPT-2, fine-tuned for theatre script generation
    - https://theaitre.com/demo
- OpenAI
  - OpenAI Playground
    - based on GPT-3, various variants
    - https://beta.openai.com/playground
  - ChatGPT
    - released 30th Nov 2022
    - based on GPT-3.5 and InstructGPT, trained with humans in the loop
    - https://chat.openai.com

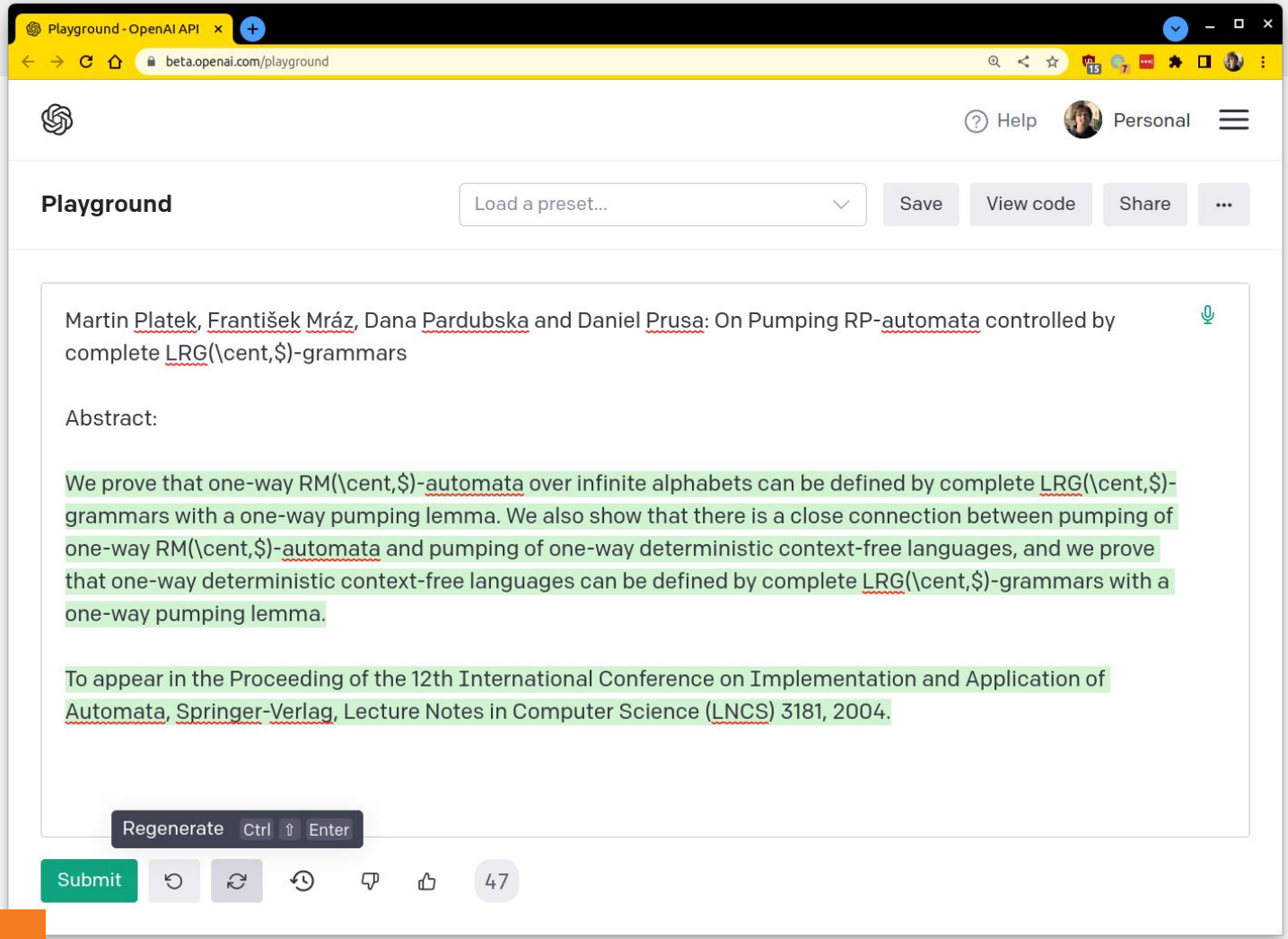

THEAITRE
THEaiTRobot 1.0
THEaiTRobot 2.0
Přepnout do češtiny
CAN A ROBOT WRITE A THEATRE PLAY?
SAMPLE SCRIPT
Scene
Polonius speaks to the king. Enter Hamlet.
Polonius
I hear him coming; let's hide, sir.
Hamlet
To be or not to be; that is the question.
GENERATE FURTHER
INPUT YOUR OWN SCRIPT
AI will generate a continuation
Scene
Write here a description of the starting situation.
Name of first character
Write here what the first character says.
Name of second character
Write here what the second character says
GENERATE FURTHER


THEaiTRobot 1.0
THEaiTRobot 2.0
Přepnout do češtiny
Scene
Helen's salon. In the room on the left, Helen plays the piano. Dominus paces the room, Dr. Gall looks out the window, and Alquist sits off to one side in a lounge chair with his face covered by his hands.
Dr. Gall
Heavens, there's more!
Domin
Robots?
Dr. Gall
That's right! Now what?
Alquist
I was afraid it would get worse.
Dominus
I am sorry, Dr. Gall. I'm afraid I have no choice but to kill you.
theaitre.com/demo

OpenAI GPT-3


Playground - OpenAI API
beta.openai.com/playground
Help
Personal
Playground
Load a preset...
Save
View code
Share
Martin Platek, František Mráz, Dana Pardubska and Daniel Prusa: On Pumping RP-automata controlled by complete LRG(\cent,$)-grammars
Abstract:
Submit Ctrl Enter
Submit
47

OpenAI GPT-3


Playground - OpenAI API
beta.openai.com/playground
Help
Personal
Playground
Load a preset...
Save
View code
Share
Martin Platek, František Mráz, Dana Pardubska and Daniel Prusa: On Pumping RP-automata controlled by complete LRG(\cent,$)-grammars
Abstract:
We prove that one-way RM(\cent,$)-automata over infinite alphabets can be defined by complete LRG(\cent,$)-grammars with a one-way pumping lemma. We also show that there is a close connection between pumping of one-way RM(\cent,$)-automata and pumping of one-way deterministic context-free languages, and we prove that one-way deterministic context-free languages can be defined by complete LRG(\cent,$)-grammars with a one-way pumping lemma.
To appear in the Proceeding of the 12th International Conference on Implementation and Application of Automata, Springer-Verlag, Lecture Notes in Computer Science (LNCS) 3181, 2004.
Regenerate Ctrl ⇧ Enter
Submit
47

ChatGPT

Please write the abstract for the paper "Analyzing Item Popularity Bias of Music Recommender Systems: Are Different Genders Equally Affected?"

I'm sorry, I couldn't find a paper with that exact title. Recommender systems research is a broad field, and there could be papers with similar titles.
It is also possible that the paper you are looking for hasn't been published yet, or that it is not publicly available.

However, a possible abstract for a paper with that title could be:

"Music recommender systems have become an integral part of many music streaming platforms, yet they have been shown to suffer from various biases, such as item popularity bias. This bias leads to the recommendation of popular items more frequently than less popular items. In this paper, we analyze the extent to which item popularity bias affects the recommendations made by music recommender systems for different genders. Using a dataset of music listening history, we compare the popularity of items recommended to male and female users and investigate whether there are significant differences in the level of bias. Our results show that both male and female users are affected by item popularity bias, but to different degrees. These findings highlight the importance of considering and addressing bias in music recommender systems, particularly in terms of gender."

This is a hypothetical abstract for a paper that might be out there, and I apologize if this is not the one you were looking for.

ChatGP

I'm sorry, I couldn't find a paper with that exact title. Recommender systems research is a broad field, and there could be papers with similar titles.
It is also possible that the paper you are looking for hasn't been published yet, or that it is not publicly available.

However, a possible abstract for a paper with that title could be:

"Music recommender systems have become an integral part of many music streaming platforms, yet they have been shown to suffer from various biases, such as item popularity bias. This bias leads to the recommendation of popular items more frequently than less popular items. In this paper, we analyze the extent to which item popularity bias affects the recommendations made by music recommender systems for different genders. Using a dataset of music listening history, we compare the popularity of items recommended to male and female users and investigate whether there are significant differences in the level of bias. Our results show that both male and female users are affected by item popularity bias, but to different degrees. These findings highlight the importance of considering and addressing bias in music recommender systems, particularly in terms of gender."

This is a hypothetical abstract for a paper that might be out there, and I apologize if this is not the one you were looking for.

Some language model limitations

- Based on available data
  - good for frequent text types, bad for rare types/topics/events/styles/genres…
  - biases, stereotypes, racism, sexism, clichés…
  - data hungry, needs very large training data (e.g. compared to people)
- Lacks knowledge of knowledge
  - intermixed language knowledge (probable text) and factual knowledge (true facts)
  - limited introspection/confidence estimation
  - research on interpretability and explainability
- Limited channels
  - text, limited multimodality, limited real-world experience, lacks human experience
- Technical limitations
  - Limited context … GPT-4: window of up to 32k tokens (dozens of pages)
  - Bad notion of time (sees texts from all time points at once, no metadata)
  - Bad representation of numbers and maths
- Understanding language? (General) Artificial Intelligence?
  - ?

Rudolf Rosa

GPT et al.: Generating Texts with Transformer-Based LLMs

- Machine Translation, Language Modelling
- RNN, Attention, Transformer
- Large Language Models

These slides available at bit.ly/theaitre-sui

https://www.theaitre.com/

The project TL03000348 THEaiTRE: Umělá inteligence autorem divadelní hry is co-financed with the state support of Technological Agency of the Czech Republic within the ÉTA 3 Programme.